

AD-A251 320



2

**FINAL REPORT
TO
OFFICE OF NAVAL RESEARCH
ON
PRECISION ENGINEERING**

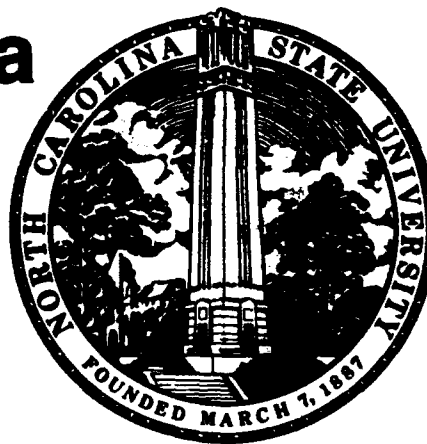
Avail:	A
Dist	A-1

**DTIC
ELECTE
JUN 1 1992**
S C D

**North Carolina
State
University**



**Precision
Engineering
Center**



DISTRIBUTION STATEMENT A

**Approved for public release;
Distribution Unlimited**

92-08698



92 4 03 220

Statement A per telecon
Dr. Ralph Wachter ONR/Code 1133
Arlington, VA 22217-5000

NWW 6/1/92



**FINAL REPORT
TO
OFFICE OF NAVAL RESEARCH
ON
PRECISION ENGINEERING**

**University Research Initiative
North Carolina State University**

N-00014-86-k-0861

Covering the period from October 1, 1986 - September 30, 1991

Accession For	
NTIS GPO	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

Faculty:

Thomas A. Dow, Editor
Robert Fornaro
Michael Paesler
Paul I. Ro
Phillip E. Russell
John Strenkowski
Ronald O. Scattergood
John M. Mackenzie, Jr.

Graduate Students:

Jeff Abler	Bill Allen	Jim Cuttino
Robert Day	Joe Drescher	Andre Fredette
Gary Hiatt	Peter Hubbel	William Larson
Michele Miller	Charles Mooney	Patrick Moyer
Walter Rosenberger	Stan Smith	Dwayne Sorrell
John Thornton	Michael Tidwell	

Staff:

Sally Bierce
Kenneth P. Garrard
Dieter P. Griffis
George Moorefield, II
Lauren Taylor
Leigh Ann Weathers
Li Zhou

TABLE OF CONTENTS

SUMMARY

i

MEASUREMENT AND MOTION

1. The Effect of Atomic Force Microscope Parameters on Image Interpretation
by *J. Thornton and P. Russell* 1
2. Scanning Optical Microscopy
by *P. Moyer and M. Paesler* 3
3. Tip Fabrication Techniques for Scanned Probe Microscopy
by *L. Zhou, H. Ximen, C. Mooney, and P. Russell* 17
4. Process Automation for Production of Controlled Geometry STM Tips
by *J. Robb, D. Griffis, J. Mackenzie, and P. Russell* 53
5. Linear Slide Actuators for Long Range Motion with Nanometer Accuracy
by *J. Cuttino and T. Dow* 63
6. DTM Metrology
by *P. Hubbel, D. Moorefield, and T. Dow* 83

FABRICATION TECHNIQUES

7. Tool Force and Surface Finish Aspects in Diamond Turning of Ductile Metals
by *J. Drescher and T. Dow* 109
8. Diamond Tool Wear
by *W. Larson and J. Strenkowski* 139
9. The Interaction of Machining Parameters on the Fracture of Brittle Materials
During Machining
by *G. Hiatt and J. Strenkowski* 157
10. Ductile-Regime Machining of Germanium
by *M. Tidwell and R. Scattergood* 175
11. Contour Grinding of Brittle Materials
by *W. Rosenberger, D. Moorefield, D. Dickey, and T. Dow* 189
12. Material Factors in Precision Machining
by *S. Smith and R. Scattergood* 203
13. Nanofabrication
by *R. Day and P. Russell* 221

MACHINE CONTROL

14. Model Reference Adaptive Control of Dual-Mode Micro/Macro Dynamics
of Ball Screws for Nanometer Motion
by P. Hubbel and P. Ro 225
15. Control of Precision Slide Motion for Vibration Reduction
by J. Abler and P. Ro 247
16. Enhancements to Three Axis DTM Controller
by M. Miller and T. Dow 265
17. Diamond Turning Machine Controller Software Development
by K. Garrard and R. Fornaro 279
18. H²ART-Based Integrated Fast Tool Servo Controller
by L. Taylor and R. Fornaro 291
19. Performance of Interprocessor Communications Architectures
by B. Allen and R. Fornaro 301
20. Performance Measurements of UNIX as a Real-Time Operating System
by D. Sorrell and R. Fornaro 319
21. An Approach to the Design and Analysis of Real-Time Computer
Control Software
by A. Fredette and R. Fornaro 333

PERSONNEL 355

ACADEMIC PROGRAM 373

SUMMARY

The focus of the research effort in the Precision Engineering Center is to improve the capability of precision manufacturing process. This focus requires research into new sensors and measurement techniques for surface characterization, in-depth studies of fabrication techniques, and hardware and software solutions for enhanced machine control.

This Annual Report summarizes the progress during 1991 of the 18 graduate students, 2 post-doctoral fellows, 4 person technical staff and 9 faculty members in the Center. The faculty and students are from 7 departments in the Colleges of Agriculture and Life Sciences, Engineering, and Physical and Mathematical Science including: Computer Science, Electrical and Computer Engineering, Material Science and Engineering, Mechanical and Aerospace Engineering, Microbiology, Physics, and Statistics.

The projects have been written as separate Sections but the interaction between projects is the strength of this organization. The multidisciplinary viewpoint provided by this unique group of faculty and students with their different backgrounds and experiences is the essence of the Precision Engineering Center. The product of the Center is new technology as well as a new cadre of researchers. The students involved in the Center, described in the back of this report, are of the highest quality.

The project summaries are arranged under three broad categories:

- Measurement and Motion
- Fabrication Techniques
- Machine Control.

MEASUREMENT AND MOTION

The emphasis in these projects has been to develop techniques that can be used to characterize fabricated surfaces at nanometer resolution as well as to study drive systems that have the potential for nanometer positioning resolution.

Atomic Force Microscopy - The AFM has become an increasingly popular tool for material characterization. It allows the imaging of a surface with a resolution as high as the atomic limit. However, the interpretation of the images is influenced by the parameters of the microscope and the tip. For example, the choice of imaging mode - constant force height or differential force mapping - may have a significant impact on the resulting image as will the relative radius of the tip and the surface features.

Scanning Optical Microscopy - The STM obtains topographic information from electrical current that tunnels between the conducting tip and the surface. In like manner, optical techniques can be used to measure a surface shape using a variety of feedback mechanisms. If light can be coupled into a surface so that it totally internally reflects, the evanescent field can be used in the same manner as the electrical field in the STM. However, this technique is limited to specific samples and shapes. A more useful technique is Near-Field Scanning Optical Microscopy (NSOM) which can be used for a wide range of sample types. Light is projected on the surface through a small diameter fiber and the reflected or transmitted light can be used to characterize not only the shape of the surface, but also important structural and chemical properties.

Tip Fabrication - One of the key components in each of the scanning microscopy techniques is the tip. This is equally true for the optical, force and tunnelling microscopes. Much effort has been devoted to developing reproducible techniques for tip preparation and also methods that can produce specific geometries for certain applications.

Nanometer Motion - Drive systems capable of nanometer positioning resolution are needed for measuring as well as manufacturing precision components. The characteristics of a ball-screw drive are defined and preliminary measurements on a traction drive are presented.

DTM Metrology - Measurements of the error motions of the ASG 2500 DTM have continued over the last year. Axis positioning errors and air bearing spindle growth have been characterized and the source of these errors have been described.

FABRICATION TECHNIQUES

Understanding the details of a fabrication process and the relationships between the operating conditions, tool geometry, and material response is the first prerequisite to control it. The processes of interest are those techniques capable of producing specular, damage free surfaces on ductile and brittle materials, for example, diamond turning and grinding.

Tool Force, Tool Wear, and Surface Finish - There are several ongoing projects attempting to define the connection between the measured tool forces, the resulting tool wear and the surface finish produced. One is principally experimental in scope where measurements of the tool forces are related to the measured edge condition of the tool as well as the surface finish of the machined ductile metal part. The second is analytical in nature using finite element models that include thermal effects to predict the wear mechanism and the magnitude of the changes in tool geometry. The third attempts to predict the onset of fracture damage in brittle materials using finite element models with corroboration via experimental measurements of surface fracture in machined brittle samples.

Ductile Regime Machining - The machine and material parameters crucial to the generation of damage free surfaces in brittle materials are being defined. Measurements of interrupted cutting tests have led to a model of brittle fabrication where material is being removed simultaneously by ductile and brittle processes. Damage free surfaces can be produced by this process if the brittle fracture does not penetrate through to the finished surface. Defining the appropriate material properties, developing techniques for measuring those properties, understanding the effect of machine and geometry variables, and corroborating those predictions with diamond turning and grinding experiments are the goals of these research efforts.

Nano-Fabrication - Recent advances in methods for imaging and manipulating materials at the nanometer scale make it possible to consider a broad array of new approaches for fabricating ultrasmall devices. Coupling carefully chosen experiments with atomic scale modeling, efforts are proposed to explore the limits of size and control achievable with new nanofabrication techniques.

MACHINE CONTROL

To obtain the advantages of real-time control of a fabrication process, the machine controller must be fast, flexible, and capable of high-speed I/O. The issues related to machine control involve defining the pertinent characteristics of the machine, developing effective control algorithms, designing multi-microprocessor architectures appropriate to the controller demands, and implementing them in software and hardware.

DTM Control - Defining the response of the drive system on a commercial DTM and improving the performance of this machine through innovative closed-loop control techniques has been the emphasis of two projects described in this report. Other enhancements to the machine tool controller of the DTM have included error correction for repeatable slideway straightness errors, and implementation of a integrated fast tool servo controller for fabricating non-rotationally symmetric surfaces.

Multiprocessor Computer Systems - The control algorithms needed for typical precision engineering applications (DTM, STM, AFM, NSOM) are similar enough that a generic solution is possible. The H²ART system architecture resolves important real-time control problems inherent in precision engineering applications. Several generations of hardware and software have been built and used for both the fabrication and measurement applications as well as basic computer science studies. The flexibility of such multiprocessor architectures creates new challenges in formulating algorithm decompositions, defining interprocessor communication mechanisms, and scheduling real-time deadlines.

1 THE EFFECT OF ATOMIC FORCE MICROSCOPE PARAMETERS ON IMAGE INTERPRETATION

John T. Thornton

Graduate Student

Phillip E. Russell

Professor

Materials Science and Engineering

1.1 INTRODUCTION

The atomic force microscope has become an increasingly popular tool for materials characterization in precision engineering. However, the uses of the AFM are limited by the interpretation of images resulting from its parameters. A thorough working knowledge of the influence of all working parameters with respect to the imaging system is needed to produce the best results.

1.2 PROJECT GOAL

Exploring the AFM parameters which contribute to image interpretation is the primary objective of this project. Once a greater understanding of image formation has been developed, the AFM will be used to examine precision engineering surfaces to display the results of materials processing techniques. The major factors which influence image interpretation are: force interactions between the tip and the sample, the electronics and digital signal processing of the imaging system, the tip shape, the imaging environment (air, lubricating oil, fluids), and the choice of imaging mode (constant force height or differential force mapping).

Initial imaging will be performed on three classes of materials. First, based on the previous results of Tidwell [1], diamond turning surfaces of brittle materials will be imaged to study the effect of DTM parameters on surface roughness and fracture [2,3]. Imaging will be performed on bare samples as well as through lubricating oil to reduce the formation of an oxide layer which degrades the image resolution. Careful manipulation of the tip-sample force curve will be required to assure imaging of the sample surface and not the surface of the oil. Contact mode imaging will be utilized and compared to 'adhesion mode' or meniscus or capillary mode. Second, ground surfaces will be imaged to study the same parameters on a sample with coarse topography. Scanning such rough surfaces will require a slow scan rate and an extended and durable tip to produce an accurate image. Third, atomically smooth surfaces will be studied to gain a better understanding of imaging conditions in this extreme limit of resolution.

The shape of the microtip is an important limiting parameter which can degrade the image resolution by the interference of tip aberrations. Contamination microtips grown at the Precision Engineering Center will be tested to study the effect of tip shapes on image interpretation [4].

Atomic force microscopy has enormous growth potential as a valuable tool in the nanometer scale manipulation of precision engineering materials. By performing materials characterization with AFM, new information can be gained about the mechanisms of processing techniques and the behavior of materials in general. However, effective use of this tool requires a high level of understanding of the imaging process itself, which is the basis of this project.

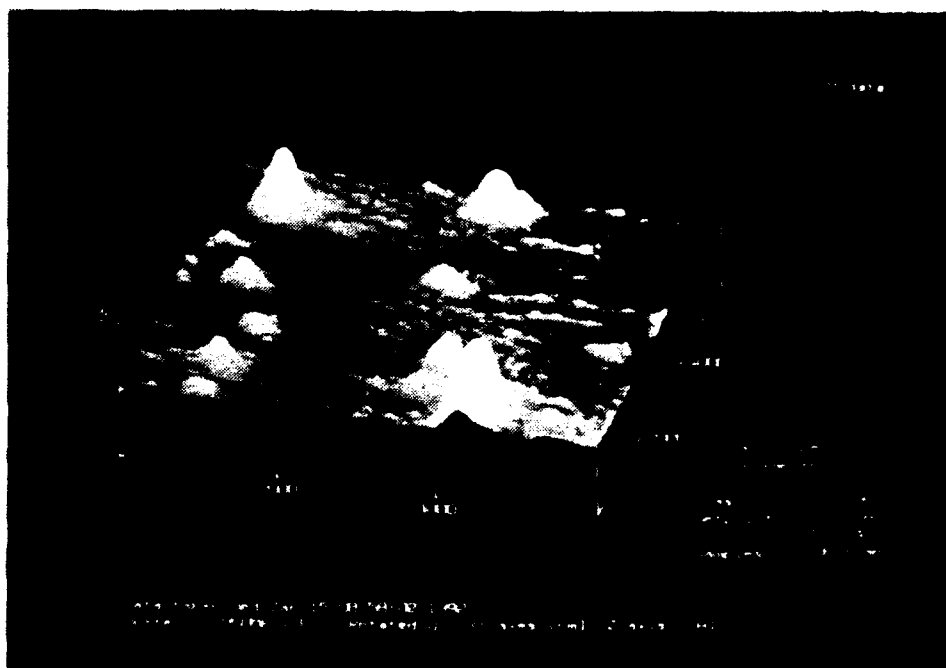
References

- [1] Tidwell, R.M., Ductile Regime Machining of Germanium: Development of New Experimental and Analytical Analysis Methods, *MS Thesis*, 1991, North Carolina State University.
- [2] Grigg, D.A., Mechanical Design of a Scanning Tunneling Microscope for the Observation of Machined Surfaces, *MS thesis*, 1989, North Carolina State University.
- [3] Biernacki, J. and Russell, P., Surface Microstructure of Precision Machined Ductile Materials, *Precision Engineering Center 1989 Annual Report*, Volume VII.
- [4] Zhou, L., Ximen, H., Mooney, C., Russell, P., Tip Fabrication Techniques for Scanning Probe Microscopy, *Precision Engineering Center 1989 Annual Report*, Volume IX.

2 SCANNING OPTICAL MICROSCOPY

Patrick J. Moyer
Graduate Student
Michael Paesler
Professor
Department of Physics

In an effort to strengthen microscopic capabilities and to broaden the range of materials that can be investigated, the optics program has been expanded to include near-field scanning optical microscopy (NSOM) as a complement to existing photon scanning tunneling microscopy (PSTM). A brief introduction to NSOM is provided. Experimental progress includes: advances in tip fabrication; the demonstration of NSOM images; and the development of a novel technique that enhances image contrast. The latter, which involves rapid modulating of the microscope sensing tip in a direction perpendicular to the surface plane of the sample, also yields information about tip structure and the local electric field distribution.



2.1 INTRODUCTION

There are two techniques for optical imaging with resolution that is not diffraction limited: near-field scanning optical microscopy (NSOM) [1] and photon scanning tunneling microscopy (PSTM) [2]. Previous reports [3] have outlined and demonstrated the capabilities of PSTM. However, there are some limitations to this technique that are not inherent in certain NSOM configurations. The PSTM exploits an exponentially decaying electric field above the surface of a sample eliminated by a beam undergoing total internal reflection. Thus these microscopes are limited to specific sample geometries as well as to optically transparent samples. These constraints do not necessarily exist in a near field scanning optical microscope.

For most precision engineering and device applications, one particular microscope configuration, reflection NSOM, provides considerable versatility in terms of the variety of samples it can accommodate. The use of the term reflective implies that the subwavelength aperture and the light delivery or collection optics are on the same side of the sample. There is nothing inherent in such a configuration which precludes the measurement of important spectroscopic signals to have value in precision engineering studies including Raman spectroscopy and photoluminescence.

To demonstrate a capability to perform NSOM experiments, PSTM instruments were adapted and transmission illumination NSOM experiments were undertaken. Results indicate that signal-to-noise is orders of magnitude greater than that observed in PSTM experiments.

2.2 NEAR-FIELD SCANNING OPTICAL MICROSCOPY (NSOM)

2.2.1 NSOM Modes

NSOM may be performed in a number of configurations each of which takes advantage of different properties of a localized light beam. Figure 1 is a schematic diagram of the geometry of the various NSOM modes. This particular schematic is modeled after an outline provided in the literature [4]. The modes of operation, as labeled, illustrate illumination, collection, and reflection mode NSOM.

In the illumination mode, incident light is coupled into the end of the fiber opposite the tip and the tip is used to deliver the light, or illuminate the surface. Using a probe which is an aperture in a conducting screen, it can be shown that in a region very close to the back side of the aperture, the light is highly collimated before diverging [5]. Thus, if such an aperture were within this collimation region (which is of the same order of magnitude as the aperture diameter) only a small region of the sample is illuminated. The light is collected on the back side of the sample with a microscope objective and focused onto a detector. The sample or tip is raster scanned to obtain an image. The spatial resolution is a function of the size of the illuminated region since it is the optical

properties of this region that affect the behavior of the detected signal. When an optical fiber is used to deliver the light, high spatial localization is obtained through the use of an optical fiber drawn (or etched) to a tip with a diameter on the order of 50-100 nm diameter. The tip is generally coated with a metal film. The details of this method appear below.

Near Field Scanning Optical Microscopy (NSOM)

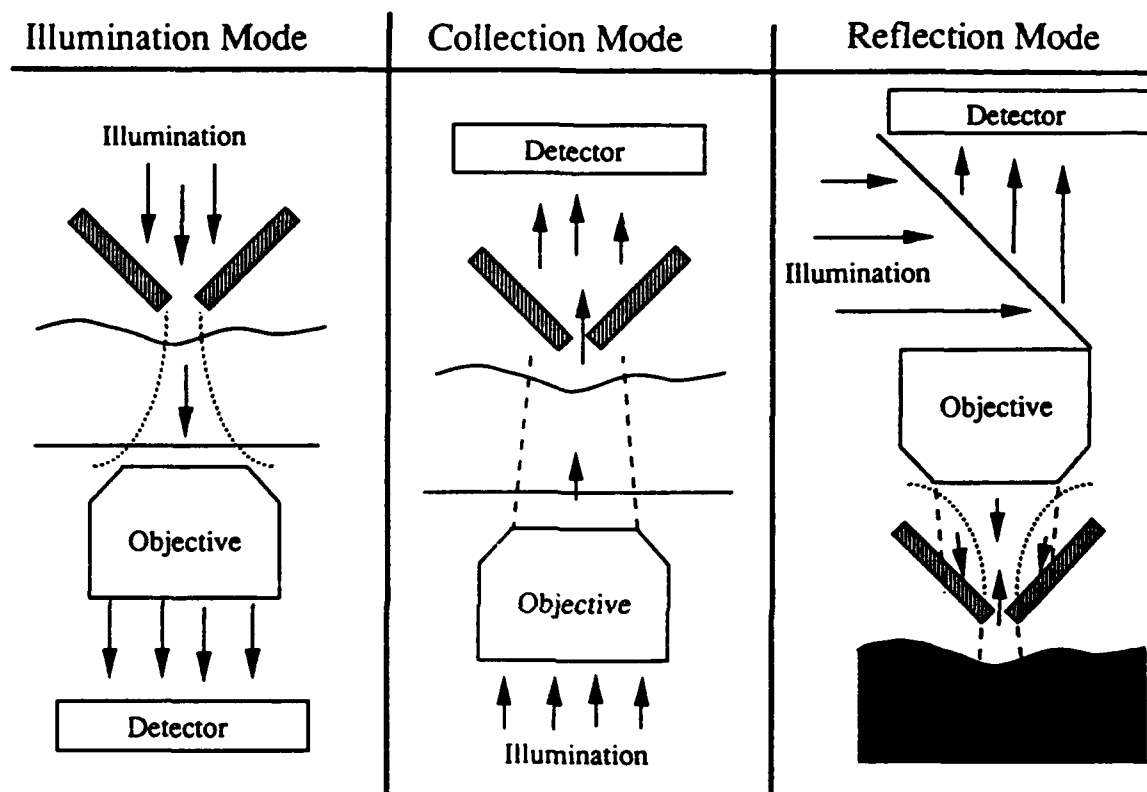


Figure 1: Schematic diagram indicating the different NSOM modes of operation.

In the collection mode, the light is delivered to the surface through a microscope objective, focused onto the surface, and then collected with an aperture or tip. The sample or tip is scanned to provide the image. The resolution in this case is of a different origin. The physical process which results in the light emerging from the sample surface may be viewed in terms of a re-emission of the light incident on the sample. If, for simplicity, one assumes the emissions are assumed to be from oscillating dipoles, the light in terms of near, intermediate, and far field electric field terms can be analyzed [6]. These fields decay as $1/r^3$, $1/r^2$, and $1/r$ respectively. However, since only the $1/r$ terms contribute to propagating energy, the other terms are non-propagating, or evanescent. That is, they are confined to the surface. Thus, if the small probe is scanned in close proximity to the surface, these fields can excite propagating fields in the fiber and be detected. In a sample for which there are strong spatial variation of these fields, resolution on the order of the tip size can be achieved.

The reflection mode has been discussed in terms of in three distinct configurations. First, the light can be delivered and detected with the sub-wavelength aperture or tip. Second, the light can be delivered with the aperture and collected with a long working distance objective above the tip and sample. A third design involves an obliquely incident laser beam focussed to a spot just under the tip. In this configuration, the reflected and/or scattered light from the surface is collected by the tip.

2.3 EXPERIMENTAL RESULTS

2.3.1 Tip Fabrication and Considerations

It should be evident that the resolution of any PSTM or NSOM is primarily dependent on tip shape and size. The resolution in the collection mode experiments (PSTM or collection NSOM) is a function of the collection optics of the microscope. The collected signal may be viewed in terms of an overlap integral involving the products of the electric fields that exist just outside the probe and those inherent to light in the probe. The probe may be considered to be a distorted waveguide, and the electric fields allowed in waveguides are well characterized. A perturbation to the conventional theory to account for non-cylindrical structures has been applied to this problem elsewhere [7]. It is fairly straightforward to determine the allowable modes of the fiber. However, it is more complicated to determine the nature of the fields immediately above the surface of the sample. These fields are due to scattering and near-field diffraction at the surface. In this sense, it is more difficult to analyze collection NSOM images than those of illumination NSOM. The mode structure is also important in illumination NSOM in that it determines the distribution of light onto the surface. It was shown in the previous report [3] that a metal coating was necessary to confine this mode for use in PSTM. For the purpose of a localized illumination pattern for NSOM, a metallic coating is also necessary.

The tip is drawn from a single-mode quartz fiber. After stripping the plastic coating, the fiber is placed in a modified capillary tube puller conventionally used in biological laboratories. The adapted instrument (made by Sutter Instruments) uses infra-red radiation from a CO₂ laser as a source of heat. Quartz has an absorption band near the energy of the laser so that by focussing the laser to a small spot on the fiber, considerable local heating is possible. A number of controllable pulling parameters affect the tip shape. The laser intensity determines the heat flux and subsequently the temperature of the glass. The beam can be scanned at various amplitudes and frequencies across the fiber or not at all. For the smallest tips possible, it is often best not to scan the beam. Tip shape (i.e., taper and size) is a function of scanning parameters. A 'velocity' parameter allows the puller to actually monitor the pulling strength through control of the speed with which the puller separates the two parts of fiber away from each other out of the region of intense local heating. Another parameter allows the time after which the laser is turned off relative to the engagement of the strong pull to be controlled to within ± 1 millisecond. The final control

parameter involves the strength of the strong pull which separates the two segments of fiber. Through control of all of these parameters, tip size from about 70 nm diameter to several microns may be drawn. In addition, the angular taper of the tip can be controlled. As an example of the type of tip used for high resolution measurements, Figure 2 shows SEM micrographs of the overall tip shape as well as the tip size. The tip diameter in this case is approximately 100 nm. This size and shape of tip shape are quite reproducible.

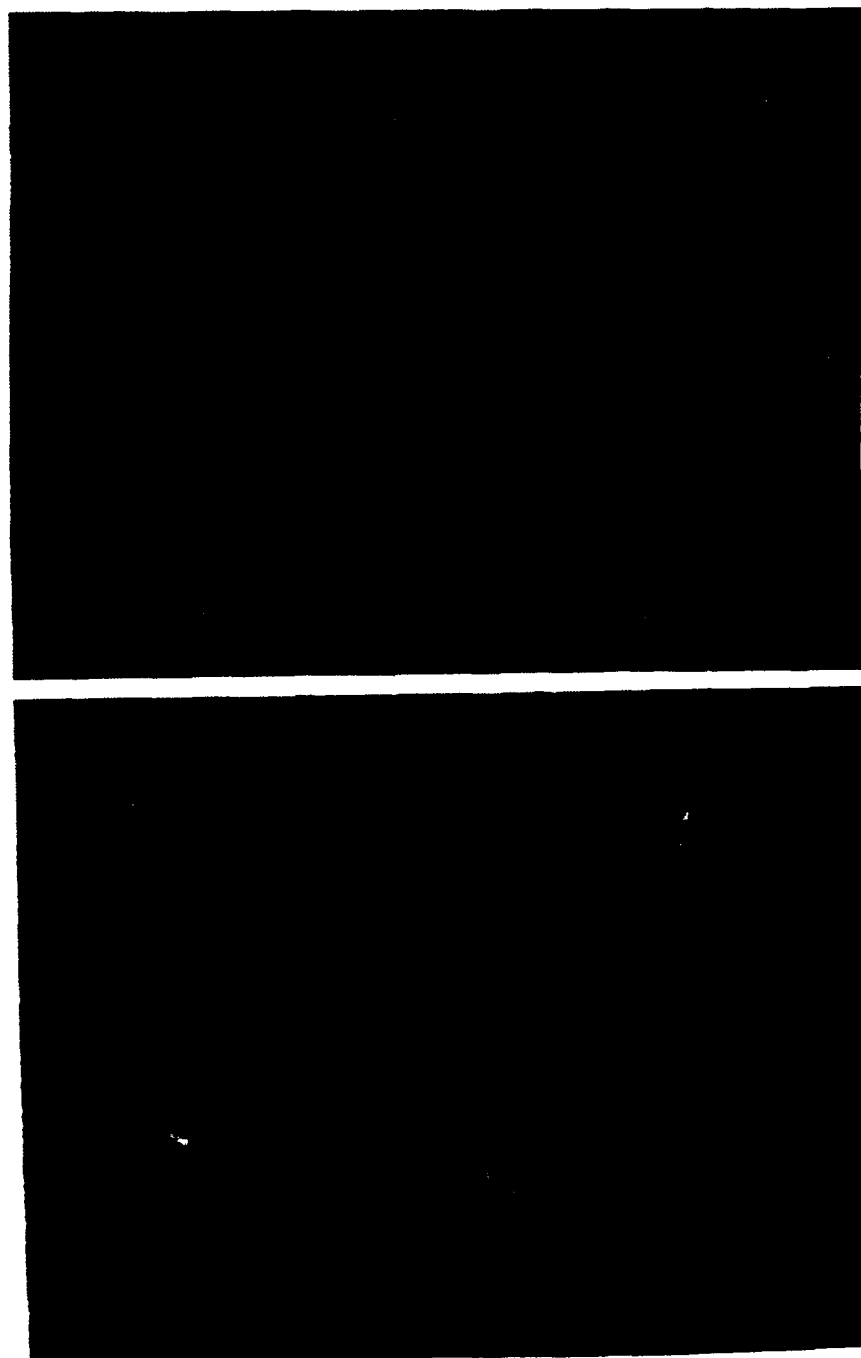


Figure 2: SEM micrographs of CO₂ pulled single-mode optical fibers. Scales are given on the images

The coating of the tips is done in a vacuum evaporator. Each tip is supported at an angle with respect to the evaporation direction. This angle is typically downward at 5-20° from horizontal. Thus the tip can be coated so as to leave a small opening at the very end of the fiber for signal transduction (collection NSOM or PSTM mode) or sample illumination (illumination NSOM mode). Stationary tips are coated with 200 to 700 Å of Aluminum (Al) and are then rotated 90 degrees and coated again. This process is iterated twice more until four evaporations cover the sides of the fiber tip. A vacuum tip rotator is currently being built so that a number of fibers can be coated in one evaporation.

2.3.2 NSOM Images

NSOM experiments were performed using an adapted Nanoscope D head fitted with a drawn optical fiber tip replacing the nominal STM tip. An air-cooled Ar⁺ laser of wavelength 514.5 nm is focused onto the non-drawn end of the fiber with a 60X objective. The base of the Nanoscope was replaced with a PEC-designed structure that allows operation in the NSOM transmission collection configuration. A microscope objective is placed below the sample and the light passing through it is detected with a silicon photodiode. The abundant signal collected with this configuration may be smaller in other configurations, so that a collector with a much lower Noise Equivalent Power (NEP) and higher quantum efficiency and gain may eventually become necessary. A photomultiplier tube (PMT) would suffice while a water cooled PMT would be ideal. Some shot noise - due to the present detector - is evident in the images to follow. Planned replacement of the diode by a water-cooled photomultiplier will reduce this problem. However the signal is detected and amplified, it is passed on to the control and imaging workstation of the Nanoscope II from Digital Instruments.

A conventional NSOM image gives the transmitted signal as a function of location in the x-y plane of the sample. Additional information can be gained by modulating, or dithering, the tip in the z-direction (i.e., perpendicular to the surface). Typical amplitudes of oscillation are 10-100 nm and the frequency is generally near 18 kHz (i.e. much higher than the scan rate.) The modulated signal is passed through a lock-in amplifier and the amplitude of the modulated signal is sent from the lock-in to the z-channel of the control and imaging electronics.

When in a dithering mode, the nature of the modulated signal is a function of the mode of operation of the microscope. In collection mode NSOM and PSTM, the component of the signal modulated at the dithering frequency is a measure of the spatial gradient of the electric fields sensed by the tip. More localized fields, i.e., those of higher spatial frequencies, have a larger spatial frequency gradient [8]. Thus, the output of the lock-in amplifier has a larger contribution from the region immediately below the tip and hence resolution is improved. In illumination NSOM the transmitted signal is monitored and the modulated component of this field provides an image corresponding to the spatial derivative of the contour of the optical density of the sample. However, by measuring only the AC component of this signal, the absolute value of the derivative is measured. The phase

of the signal can be used to determine the sign of the derivative. Thus, amplitude and phase information combine to give a complete derivative image. The dithering technique also serves to reduce the noise to only a narrow bandwidth about the dithering frequency - in this case 18 kHz.

Figure 3 shows a scanning electron micrograph of a grating that was used for imaging with both of the above mentioned techniques. The grating has 500 nm wide lines of Cr on a glass substrate. The line spacing is approximately 700 nm between lines. Figure 4 shows an NSOM image taken with a non-ideal tip. The grating is clearly evident. The small bright spots are due to shot noise of the photodiode. The bright stripes represent transmission through the sample where the metallic lines do not exist. The dark lines correspond to the Cr lines. A line scan is drawn across the grating and its periodicity highlighted. The spatial frequency highlighted in the frequency spectrum which corresponds to 840 nm is related to the line spacing. Due to the asymmetry between the line width and line spacing, a lower frequency component does not exist. The random peaks in the line scan are due to the detector shot noise. Figure 5 shows a modulated NSOM image of the same region imaged in Figure 4. Notice that the bright lines in this image correspond exactly to the edges of the lines in the DC NSOM image of Figure 4. The noise level is much lower and the edges are much sharper in the modulated image. In this image it is clear that the dithering technique provides an absolute value derivative image. The discontinuity in the line in the middle of the image of Figure 4 is also imaged clearly in the dithering experiment.

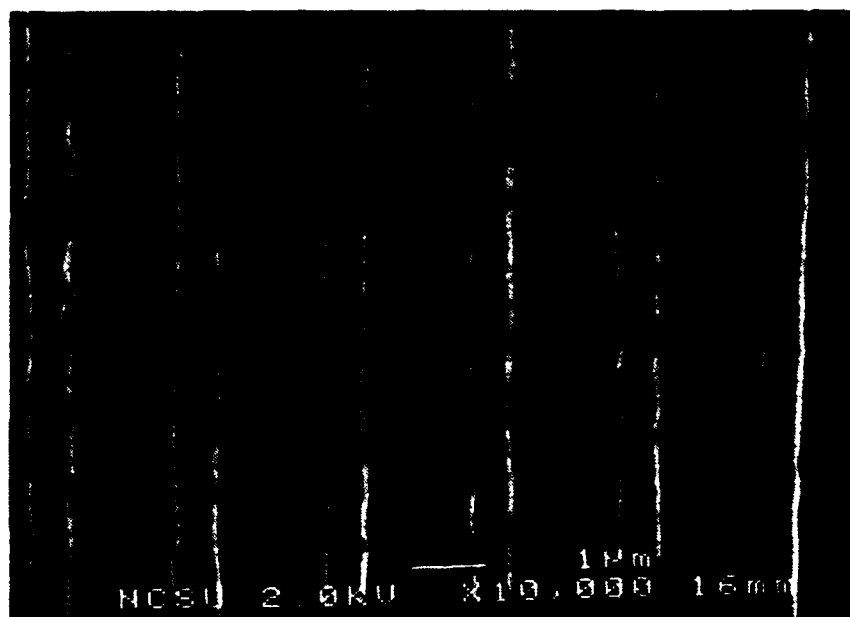


Figure 3: SEM micrograph of lined sample. Lines are deposited Cr. Scale is given on image. Line width is approximately 0.5 μm .

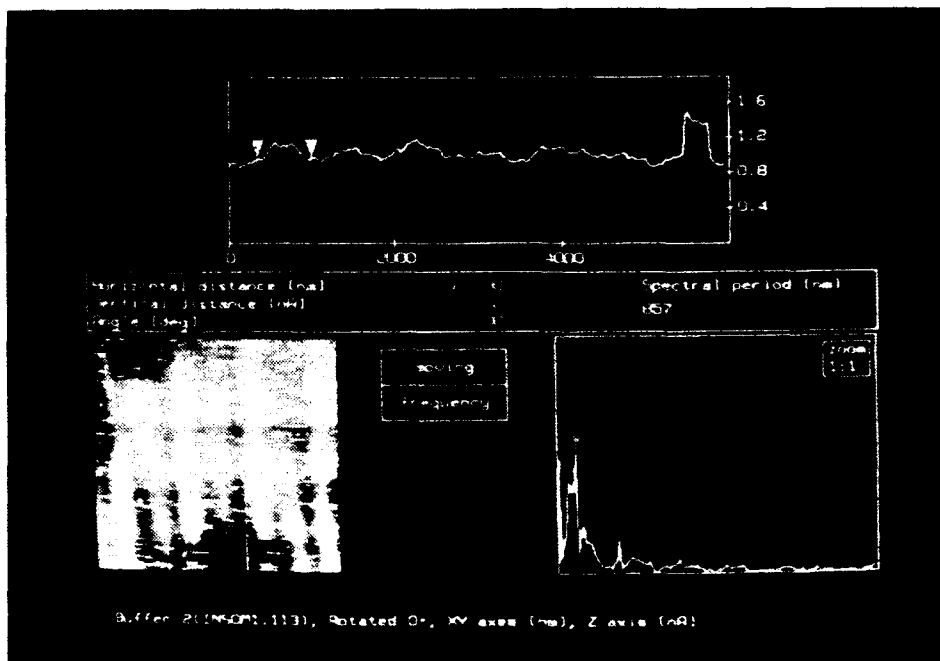


Figure 4: Illumination NSOM image of sample shown in Figure 3 (lower left). Image is $6\text{ }\mu\text{m} \times 6\text{ }\mu\text{m}$. Also shown are a line scan (top) and its corresponding frequency spectrum (lower right).

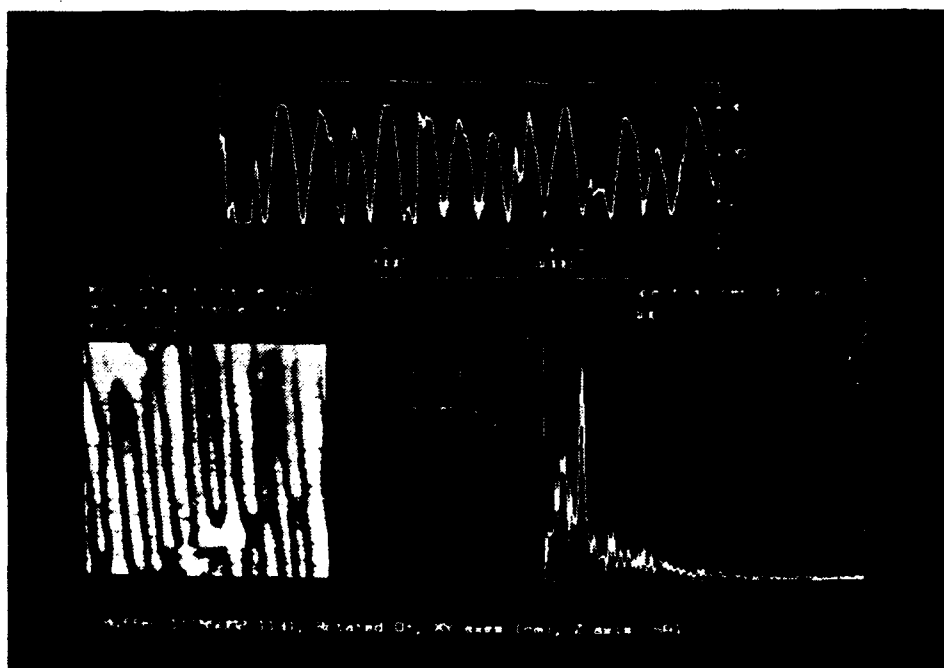


Figure 5: Modulated illumination NSOM image of sample shown in Figure 3 (lower left). Image is $6\text{ }\mu\text{m} \times 6\text{ }\mu\text{m}$. Also shown are a line scan (top) and its corresponding frequency spectrum (lower right).

Figure 6 shows an SEM micrograph of another sample. This sample is a glass substrate onto which 110 nm spheres were dispersed. The sample was then coated with approximately 150 Å of Cr. The surface of the Cr was then washed with an organic solvent to dissolve the spheres. The micrograph of this particular section of the sample shows that only a small density of holes are present relative to the density of remaining spheres. Figure 7 shows an NSOM image of the same surface. The spheres/holes appear to be about 400 nm in size. Since the SEM micrograph of Figure 3 indicates a tip diameter of about 100 nm, it appears that the metal coating is not evaporated near enough to the end of the fiber. That is, too much light is being coupled into the side of the tip. By directing the tip more in the direction of the evaporant stream during coating, this problem should be obviated.

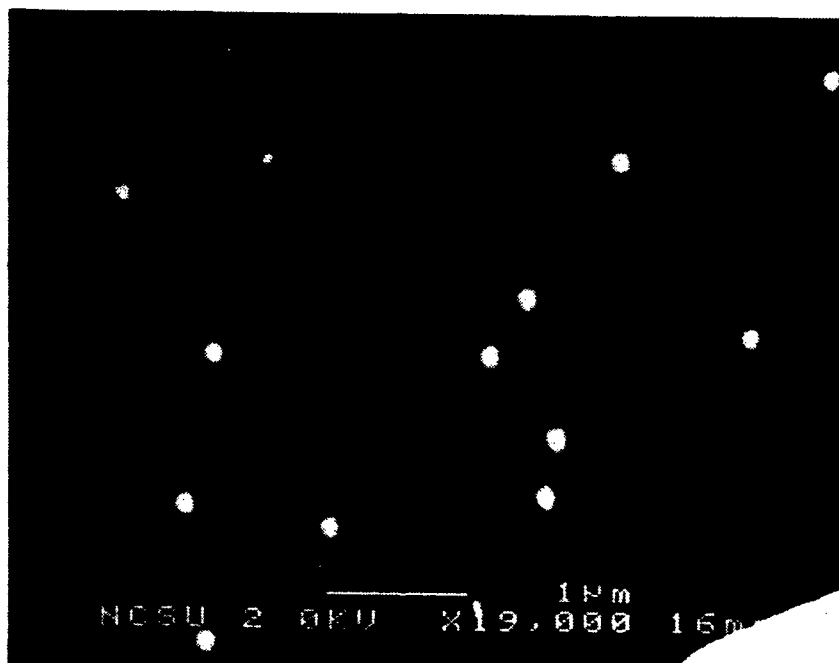


Figure 6: SEM micrograph of 110 nm spheres (bright spots) and resultant holes (dark spots) on 150 Å Cr on quartz. Scale given on image.

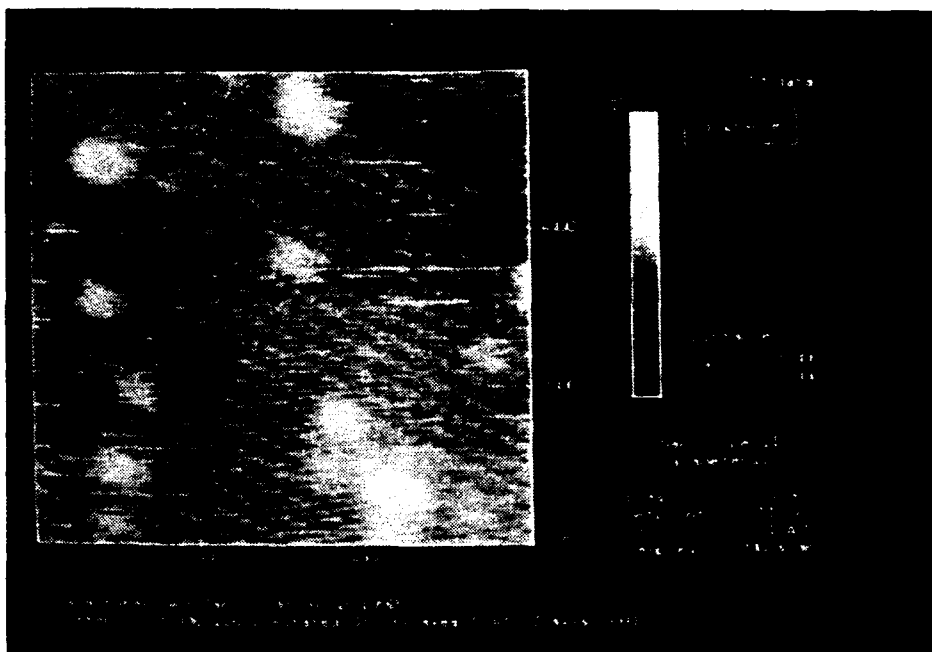


Figure 7: Illumination NSOM image of sample shown in figure 6. Image is $3\ \mu\text{m} \times 3\ \mu\text{m}$.
Texturing is caused by 60 Hz pickup.

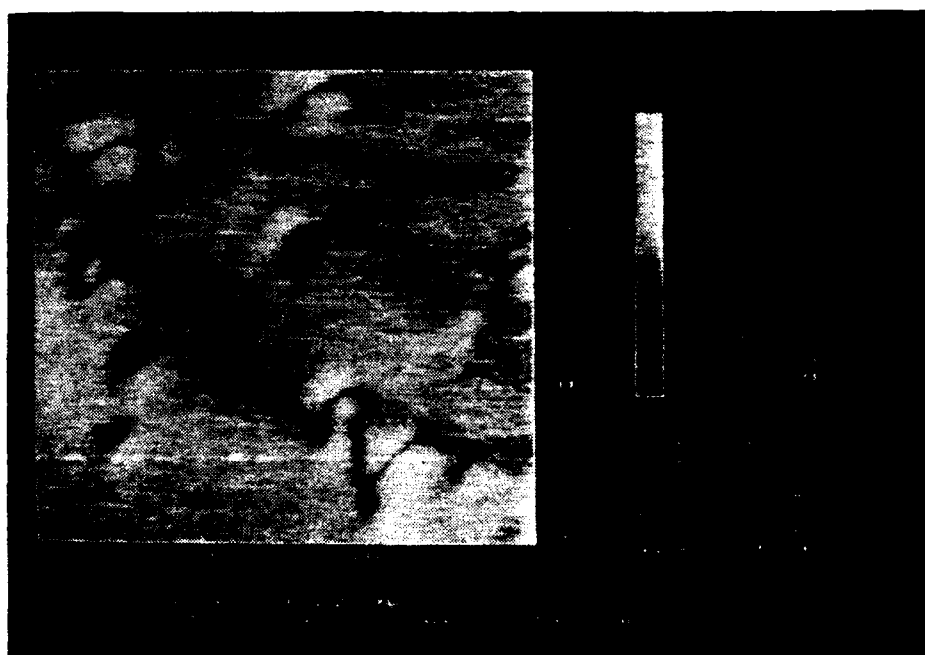


Figure 8: Modulated illumination NSOM image of sample shown in figure 6. Image is $3\ \mu\text{m} \times 3\ \mu\text{m}$.
Figure 8 is an image of the same region as shown in Figure 7 but taken with a dithered tip. If the dithering technique results in merely a signal which represents the spatial derivative of the surface profile, this image should show a bright ring corresponding to the perimeter of the hole. The distorted image of Figure 8, however, exhibits structures considerably more complicated than

Figure 8 is an image of the same region as shown in Figure 7 but taken with a dithered tip. If the dithering technique results in merely a signal which represents the spatial derivative of the surface profile, this image should show a bright ring corresponding to the perimeter of the hole. The distorted image of Figure 8, however, exhibits structures considerably more complicated than simple rings. This image can be understood in terms of an uneven metallic coating of the fiber. Apparently, a thicker coating was deposited onto the fiber on the sides where the bright spots appear. In this case, the radiation from the tip forms an elliptical profile. The degree of collimation of light emergent from the tip is a function of the thickness of the metallic coating at the tip, and the asymmetry in the coating thickness manifests itself in the non-circular images of Figure 8.

Because of the increased sensitivity of an image to tip structure observed when dithering, the throughput function of a tip can be monitored by measuring a known structure. The modulation technique can thus provide valuable information not easily extracted from conventional imaging. The imaging of holes in the thin Cr film shows that tip coating uniformity and tip size can be assessed by examining image provided with this technique. Dithering thus provides a convenient and inexpensive way to monitor tip fabrication methods in the laboratory.

2.4 FUTURE

Tip fabrication involves the drawing of a taped fiber and its subsequent coating. The drawing process is under control such that small radii tips can be reproducibly drawn. The coating process, although not completely repeatable, should improve when of the fiber rotator-coater nearing completion in the machine shop is installed. With both aspects of the tip fabrication operations under control, a careful study correlating SEM micrographs and NSOM image quality as a function of tip morphology and coating will be a high priority.

With successful illumination NSOM results in hand, adaption to reflection NSOM represents the next experimental goal. As shown above, the modifications necessary to change the existing system to reflection NSOM are not extensive. The ability to image optically opaque samples would add considerable versatility to the system. The modifications necessary for this type of microscope are currently being implemented.

One of the most significant disadvantages of NSOM is that no feedback mechanism exists for maintaining a constant tip-sample separation. The PSTM technique, although not as versatile as the various NSOM configurations, does have this advantage. Currently, it seems that the best technique available to couple to an NSOM is some form of force feedback. A dual shear force, NSOM has recently been built [9]. A schematic diagram of this experiment is shown in Figure 9. In this microscope the tip is dithered in a direction lateral to the surface. The tip is oscillated with a frequency equal to a resonance of the tip. The amplitude and phase of the tip is then monitored as a

function of distance to the surface. Ambient conditions (e.g., humidity) will affect this functional dependence. The amplitude of the lateral oscillations is monitored in the following manner: The light from the tip is focused onto a pinhole which is offset spatially in the focal plane in the direction of dithering such that the pinhole is essentially located on the maximum slope of the imaged Gaussian profile. As the tip is dithered, the profile moves and the magnitude of the AC component measured is a function of the oscillating amplitude. If a combination of the magnitude and phase of this signal is used as feedback, the tip-sample separation will remain constant. Thus, a plot of the z-voltage will provide a shear force image. The DC optical signal through the pinhole (normalized to account for changes in transmissivity) can be plotted to give an NSOM image. Plans to incorporate this dithering technique are underway.

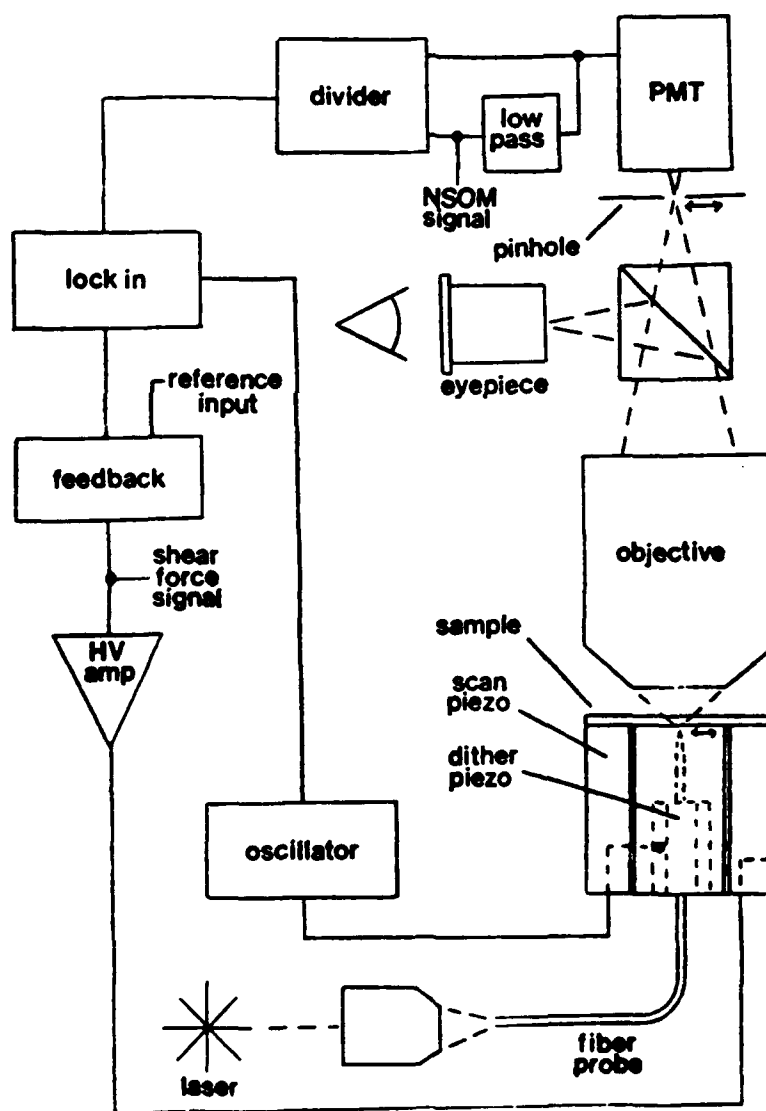


Figure 9: Schematic diagram of dual shear force microscope and NSOM under construction [9].

2.5 CONCLUSIONS

Near-field scanning optical microscopy (NSOM) has been introduced and the basic modes of operation have been described. The expansion of the optical laboratory to include the capabilities to perform such experiments has been discussed and data has been shown to indicate progress in this direction. Tip fabrication techniques have also been outlined and results shown. Finally, a novel NSOM and PSTM imaging technique has been introduced that provides optical information not typically provided by conventional imaging techniques.

References

- [1] Harootunian, E., Betzig, E., Isaacson, M., and Lewis, A., "Super-resolution fluorescence near-field scanning optical microscopy", *Appl. Phys. Lett.*, vol 49, p. 674, 1986.
- [2] Reddick, R., Warmack, R., Ferrell, T., "New form of scanning optical microscopy", *Phys. Rev. B*, vol 39, p. 767, 1989.
- [3] Moyer, P.J., "Analytical Photon Scanning Tunneling Microscopy", *Precision Engineering Center 1990 Annual Report*, NCSU, Raleigh, NC, Vol VIII, December 1990, pp. 29-46.
- [4] Betzig, E., Isaacson, M., Barshatzky, H., Lewis, A., and Lin, K., "Near-field scanning optical microscopy (NSOM)", *Proceedings of the SPIE Vol. 897 Scanning Microscopy Technologies and Applications (1988)*, p. 91.
- [5] Betzig, E., Harootunian, A., Lewis, A., and Isaacson, M., "Near-field diffraction by a slit", *Applied Optics*, vol. 25, p. 1890, 1986.
- [6] Jackson, J.D., *Classical Electrodynamics*, John Wiley and Sons, New York, 1975, p. 395.
- [7] Buckland, E., Moyer, P.J., Paesler, M.A., "Optical Coupling and Resolution Issues in Scanning Optical Microscopy", submitted to *Physical Review*.
- [8] Goodman, J.W., *Introduction to Fourier Optics*, McGraw-Hill, San Francisco, 1968, p. 51.
- [9] Betzig, E., Finn, P.L., and Weiner, J.S., "Combined Shear Force and Near-field Scanning Optical Microscopy", to be published.

3 TIP FABRICATION TECHNIQUES FOR SCANNING PROBE MICROSCOPY

Li Zhou

Post Doctoral Research Associate

Hongyu Ximen

Post Doctoral Research Associate

Chuck Mooney

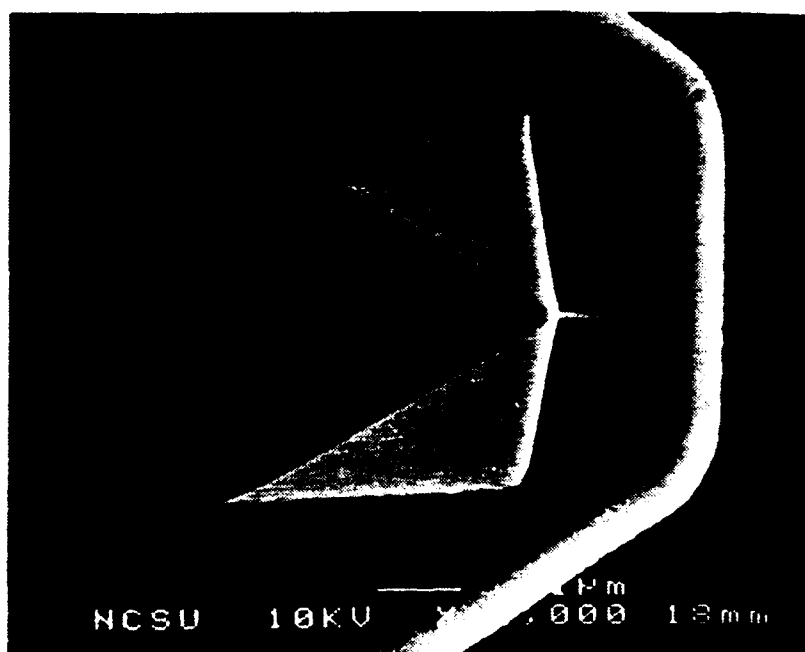
Graduate Student

Phillip E. Russell

Professor

Materials Science and Engineering

Two techniques for fabrication of microtips on conventional atomic force microscope (AFM) pyramidal tips have been developed. The Focused Ion Beam Micromachining (FIBM) technique is used to selectively remove certain areas of the pyramidal tip so that the remaining portion serves as a sharper tip. Microtips can also be made by electron beam deposition utilizing the residual gas in the vacuum chamber. The process is reproducible and the resulting tips have high rigidity and excellent durability in AFM imaging. AFM images of high aspect ratio lithographic patterns using microtips show fewer artifacts than conventional pyramidal tips. Initial progress on growth of controlled shape microtip to serve in atomic resolution measurements has been made.



3.1 INTRODUCTION

As a logical extension of the Scanning Tunneling Microscope (STM) tip fabrication projects, an investigation into improving the tip fabrication technique for Atomic Force Microcopy (AFM) tips has been undertaken. The problems addressed here relate to the artifacts that are produced from imaging high aspect ratio (depth/width) features with standard AFM pyramidal tips. The problem of what to use as a force sensing tip has existed since the invention of AFM. The first tips were simply the end of a cantilever with no pyramid. Improvements were made by either evaporating material through a small hole forming a cone or by glueing shards of diamond to a Si_3N_4 cantilever [1]. The problem with these tips was irregular geometry. That problem was solved by the development of integrated pyramidal tips (trade name Nano Probe) which are now the industry standard [2]. Standard tips still present a problem: the pyramid's base is 3 microns square and the walls are angled at 55° . This tip, while having a regular geometry, cause a great deal of "smoothing" to high aspect ratio topological features. The use of microtips eliminates many of the tip artifacts found when using standard pyramidal tips.

Two techniques have been developed, additive and subtractive. The first method attempted was removing part of the standard AFM tip by Focused Ion Beam (FIB) micromachining so that the left-over serves as a sharper tip. The second method is an additive process by which microtips are grown on the end of the force sensing pyramid. The additive process has two branches of study, direct electron beam induced contamination growth and Ga^+ ion implantation with subsequent electron beam induced growth. The first grown tips actually consisted of contamination, however, they were unreliable and results were unreproducible. The focus then shifted to modifying the existing cantilevers by Focused Ion Beam micromachining (FIBM) [3,4]. It was noticed that tips grew quite easily on areas that were Ga^+ implanted. This shifted the focus to the manufacture and testing of Ga^+ implanted pyramids with grown tips. The research was broadened to include contamination with no Ga^+ implantation. The techniques for growing straight contamination tips have now been refined and the comparison of the two is in its initial stages. Since it is still unknown whether the two additive processes are different, both have been explored on the assumption that they are not the same and thus they are discussed separately in this section of the report.

All of the grown tips, whether they are Ga^+ implanted or not, are grown in a JEOL 6400F Field-Emission SEM from background contaminants present in the imaging chamber and on the sample. At present it is unknown whether the Ga^+ implantation causes a change in the chemistry of the grown tips. This contamination comes from the polymerization of hydrocarbon vapors which backstream from the mechanical roughing pumps and diffusion pumps. It is generally believed that the most important parameters in contamination growth are current density and the amount of hydrocarbons available to polymerize. This phenomenon is well know and described in the literature [5,6,7,8,9]. Studies of contamination micro-cones [10] and self-supporting filaments

[11] have also been done in the past. The first work that used contamination cones as tips was done by Fuji et al [12]. They grew a tip on the end of a Si_3N_4 cantilever that had no pyramidal tip in order to probe a "deep structure". Other methods to fabricate microtips with electron beam irradiation have also been used [13]. The work described in this section is intended to refine microtip fabrication and understand the parameters involved. Both types of extended microtip have successfully been used to image various materials for reasonable periods of time with a Nanoscope II AFM.

3.2 IMAGING ARTIFACTS

3.2.1 Imaging Artifacts from Standard AFM Tips

Standard tips produce artifacts which can be interpreted from the geometry of the commercially available pyramidal tips. A standard AFM tip consists of a Si_3N_4 pyramid with a base that is $3 \times 3 \mu\text{m}$ with 55° walls integrated on a cantilever made of the same material. The relative size of the pyramid produces a "smoothing" effect which does not show the true topography. Since the geometry is well known, removing the tip artifacts is not difficult for surfaces which have known topographies, however, most samples that are viewed are unknown and thus do not have a known topography. For this reason it may become difficult to interpret these images. If the surface in question is relatively smooth (no features with aspect ratios of greater than 55° , or atomic resolution samples), no problems arise, however, if the surface is expected to have topography that has high aspect ratio features of size relative to the tip, then there will be induced tip artifacts such as shown in Figure 1 (a and b).

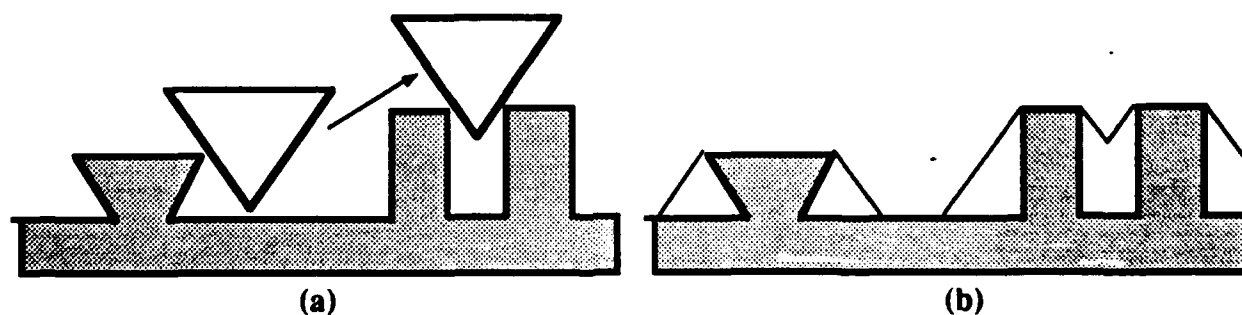


Figure 1: (a) Side view of a sample with unknown topography being imaged by a standard AFM tip. The problems with standard tips can be seen easily.
 (b) The thin lines represent the topography that would be measured with a standard AFM tip. Clearly it would be preferable to have a tip that could image between the two close towers.

3.2.2 Reduction of Artifacts with Grown Tips

Grown microtips circumvent this problem by reducing the amount of "smoothing" by creating tips with high aspect ratios that are uniformly straight (if grown properly). This allows the imaging of trenches and steep sidewalls with reduced artifacts. One must bear in mind that the image displayed does not show the true size of the topological features but the size of the features plus the size of the tip. Therefore the smaller the diameter of the tip the closer the image will be to reality.

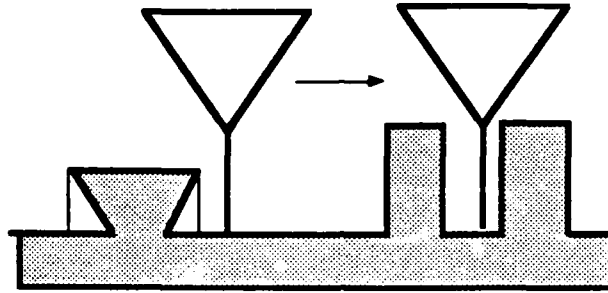


Figure 2: A microtip would give a truer representation of the topography. Unfortunately, due to limitations with the Nanoscope II, right angles are not measurable (90° sidewalls would represent two data points in the same space which is not allowed with present software).



Figure 3: Several "custom" tips grown on the side of a Si_3N_4 cantilever illustrating the degree to which the shape of the tips can be controlled.

As can be seen from Figure 2, the tip is incapable of measuring undercut sidewalls or even vertical sidewalls. Undercut walls and negative slopes may be measured with custom tips grown for a specific purpose [13]. Nyysönen, Landstein, and Coombs have developed a system to measure undercut sidewalls that does not scan but places the tip by way of computer control until the force of an approaching sidewall is measured. In this method precise dimensions of the tip must be known to subtract from the image that is extrapolated from the data [14]. Digital Instruments has also been working on a system to image undercut sidewalls. Figure 3 shows some tips that were manually grown during the first phase of contamination tip research illustrating the ability to change growth direction. Computer aided control of the electron beam should provide more control of the shape and precise dimensional control of this type of tip.

3.3 TIP FABRICATION USING THE COMBINATION OF FOCUSED ION AND ELECTRON BEAM TECHNIQUES

3.3.1 AFM Cut Tips Using Focused Ion Beam Micromachining

The FIBM has the capabilities of fabricating various types of patterns in submicron scale, such as precision cross-sectioning of integrated circuits and sharpening of STM tips [15]. A digitized three dimensional ion beam scan strategy was implemented in the focused ion beam work station such that the ion beam could be scanned in a desired pattern to micromachine commercial AFM Si_3N_4 pyramidal tips. The two basic cutting strategies are shown in Figure 4 (a and b). In the SEM micrograph of Figure 5, three sides of a pyramidal tip are shown to be cut off from the cantilever substrate. In this way, the left-over of the pyramidal tip may be able to serve as a sharper tip for AFM scan.

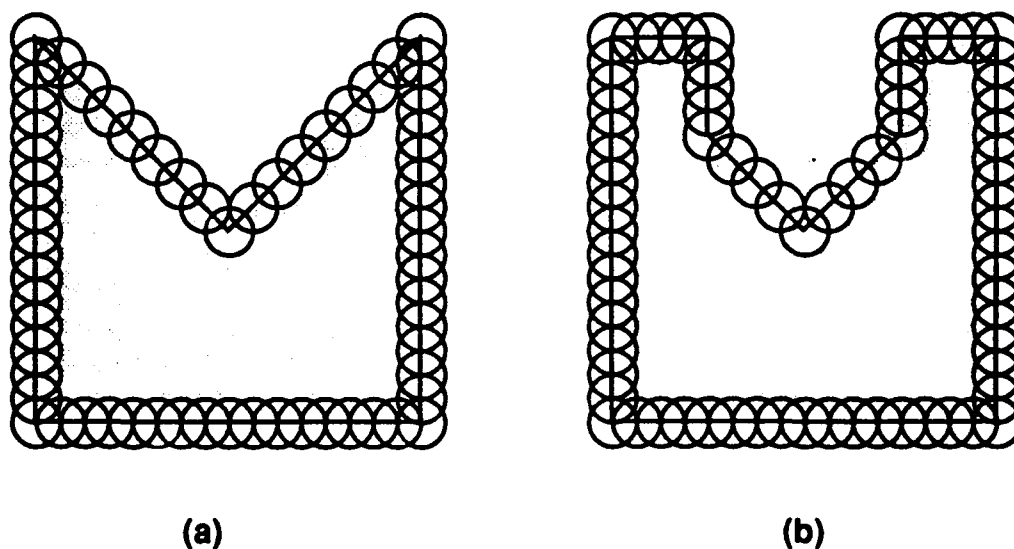


Figure 4: The FIB cutting strategies for initial and trimming cuts of AFM pyramidal tips.



Figure 5: 45 degree tilt view SEM micrograph of an AFM cut tip. The cut out section of the pyramid is still attached for clarity (It is completely detached for actual use).

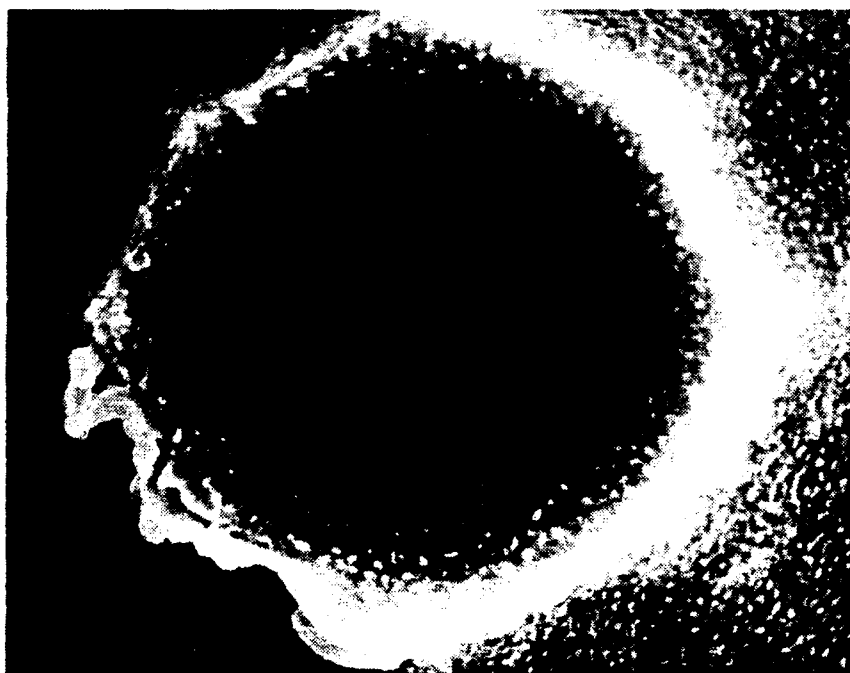


Figure 6: SEM micrograph of FIB machined scanning ion conductance microscopy (SICM) tips. A focused ion beam system with resolution of 50 nm was used in the fabrication. In this case the hole diameter is 0.2 micron.

Also, for special applications, such as scanning ion conductance microscopy (SICM), a pyramidal tip with a submicron hole on the tip apex may be desired [16]. This type of structure was fabricated by locking the focused ion beam in the spot mode, as shown in the SEM micrograph of Figure 6. The size of the hole is limited by the spot size of the ion beam. With a new focused ion beam system (FEI Co., Beaverton, Oregon) a 50 nm beam spot size is obtainable allowing a hole with of 0.2 μm diameter to be produced.

3.3.2 Electron Beam Induced Growth of Microtips on Ga^+ Ion Implanted Surface

The microtips are grown using a multistep process. The Si_3N_4 tips are first coated with gold or platinum to avoid beam charging in subsequent processing steps. The metal coating is an important step since the coated metal layer, if it is too thin, will not be continuous and prevent the surface from charging under the bombardment of a charged particle beam. On the other hand, a relatively thick layer of metal will cause the 0.7 μm thick Si_3N_4 cantilever to curl which may adversely affect AFM imaging. A cantilever that was coated with 70 \AA of gold was measured by a laser interferometer indicating the cantilever had curled 1 μm . This thickness seems to work well.

After the metal coating, the tips are placed in a focused ion beam (FIB) workstation incorporating a liquid metal ion (Ga^+) source. A Ga^+ ion beam of 250 nm diameter operating at 20 keV is used to sputter away the Au coating at the tip apex and to implant Ga ions into the tip region with a dose of $2.0 \times 10^{17}/\text{cm}^2$. This results in a local supersaturation of Ga in the pyramid area of the cantilever. For a dose of $1.0 \times 10^{16}/\text{cm}^2$, Ga in Si_3N_4 will approximately reach its peak concentration at sputter equilibrium [17]. These Si_3N_4 tips supersaturated with Ga are then transferred into a field emission scanning electron microscope (SEM) with specimen chamber pressure in the range of 10^{-7} torr. An electron beam with 30 \AA spot size is held in spot mode on the desired growth position for one to five minutes (depending on the length desired).

This modified AFM pyramidal tip with a grown microtip is shown in the SEM micrograph of Figure 7, before AFM imaging in the contact profiling mode for tens of hours. After the AFM image acquisition, the sharpness of the microtip was reduced slightly due to the contact between the microtip and sample surfaces, as seen in Figure 8.



Figure 7: SEM micrograph of an electron beam induced growth microtip grown on the top of a commercial pyramidal AFM tip which was implanted with Ga^+ ions before AFM imaging. The tip was grown for 3 minutes at the 10 kV electron beam accelerating voltage and condenser lens excitation setting of the SEM is 7 (beam current 2×10^{-11} A).



Figure 8: SEM micrograph of an electron beam induced growth microtip grown on the top of a commercial pyramidal AFM tip which was implanted with Ga^+ ions after extensive AFM (30 hours AFM scanning) imaging in the contact mode.

3.3.3 Microtip Characterization

The geometry of the microtip can be controlled by changing microscope parameters such as beam energy or accelerating voltage, condenser lens excitation and focusing properties of the electron beam as well as electron beam dwell time. It is very important to study the dependence of the tip geometry on one of these parameters while keeping the others constant because the tip geometry is very sensitive to the above parameters. The geometry, including the length, base diameter and sharpness of the tip, are essential in AFM imaging. For example, a tip that has a long and straight shank is good for imaging high aspect ratio pattern on the one hand, but may not be useful in atomic resolution measurement, due to tip flexing. Therefore, a tip with short length and good end sharpness may be necessary in the use of atomic scale AFM imaging.

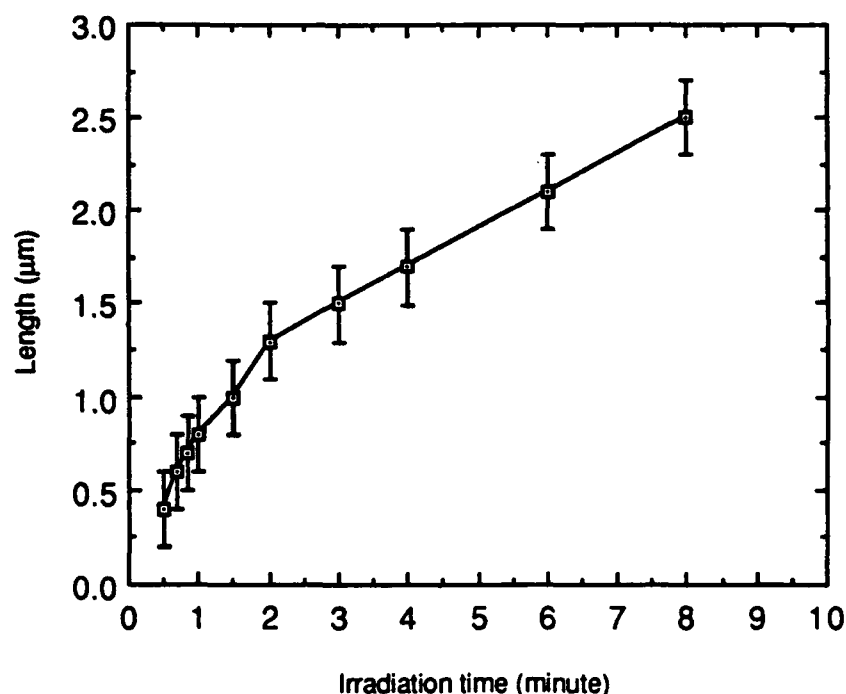


Figure 9: Measured microtip length as a function of electron beam irradiation time. The beam energy is 10 kV and condenser lens setting is 7 (2×10^{-11} A beam current).

Figure 9 shows the microtip length as a function of the electron beam dwell time, in which case the accelerating voltage is kept at 10 kV and the condenser lens excitation is 7 which means the final beam current is about 2×10^{-11} A. After an initial quick growth, the tip grows at a relatively low and constant rate when the dwell time goes beyond two minutes. Since the beam conditions are maintained fixed, the interaction between the tip and incoming electron beam causes strong electron scattering as the tip grows longer, which may result in an decrease of the growth rate. In general, the tip length could be controlled within about 10 nm depending on the electron beam stability.

Therefore, the microtip length should be controllable by the electron beam dwell time in the fine focusing condition.

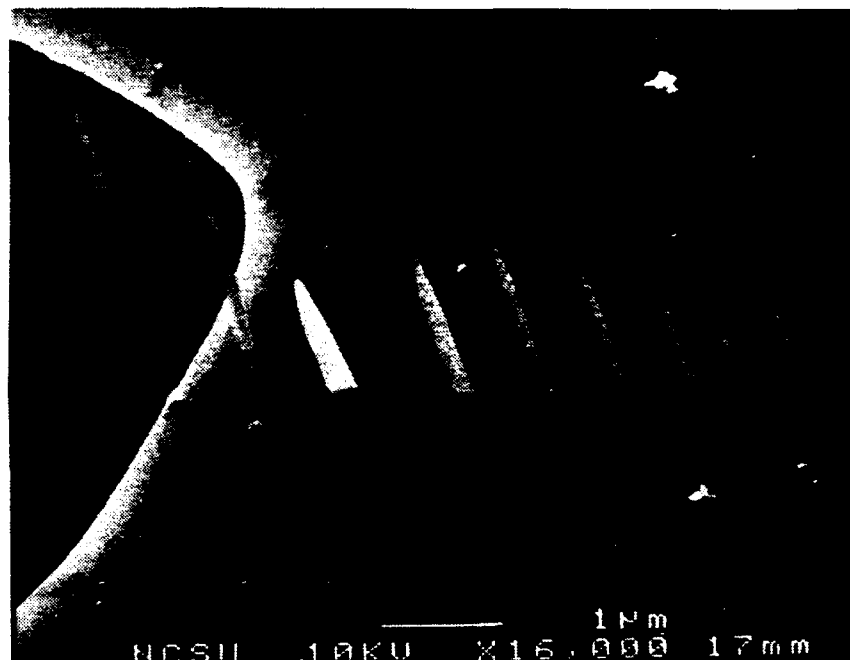


Figure 10: An SEM micrograph of an array of microtips grown at different electron beam irradiation time.

Figure 10 shows an SEM micrograph of an array of microtips grown with different electron beam irradiation times. One of the tips was inadvertently grown when the electron beam was out of focus. The tip is short and fat which demonstrates that the tip geometry changes dramatically with defocusing. Beam stability is another important issue. It is interesting to note that no matter how fine the focus is, which means that the electron beam spot size is around 30 \AA , it is impossible to make 30 \AA diameter tip. This may be due to vibration of the electron beam making the actual diameter of tip much larger than 30 \AA . As discussed in more detail in Section 1.5 where a finely controlled tip was made to use in atomic resolution measurement, it can be seen that with less the dwell time, it is possible to make tips with 20 nm diameter because the beam stability problem will be dominate less as the dwell time is shortened.

A study of the tip growth at different accelerating voltages confirmed the concept that the electron scattering plays an important role on the growth rate. Figures 11 and 12 show the microtips produced under 2 kV and 15 kV accelerating voltages, respectively. The tip grown at 2 kV has a longer shank and a dull end apex, and the tip grown at 15 kV has a shorter shank and a sharp end. The Condenser Lens (CL) excitation setting of the SEM is another important parameter in the tip fabrication. Not only the final current delivered to the target, but also the resolution of the electron beam changes significantly with condenser lens. By choosing different CL setting, the electron beam current density is actually changed. Figures 13 and 14 show significant changes in the tip length and end sharpness with different CL excitation.

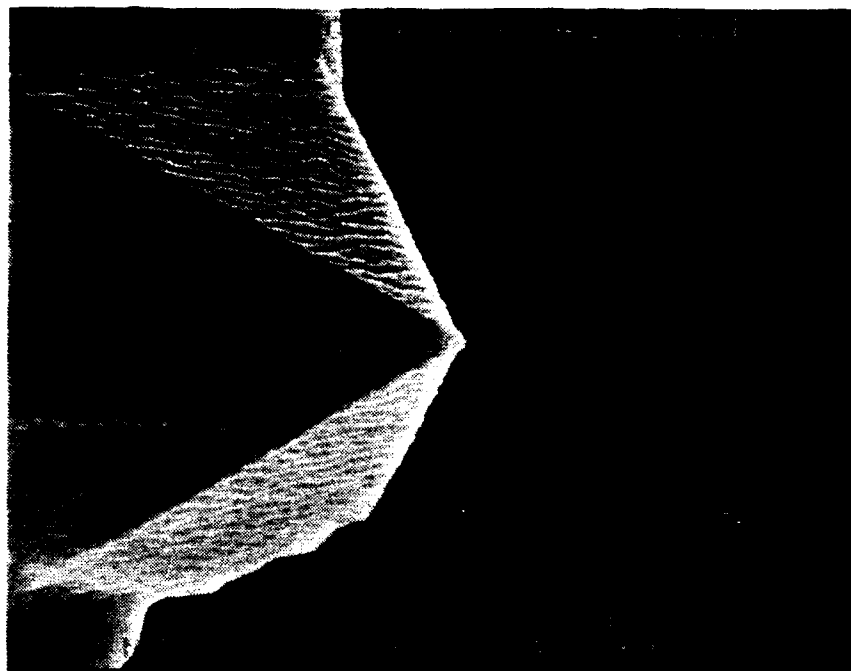


Figure 11: SEM micrograph of a microtip grown at an accelerating voltage of 2 kV. The tip is grown for 3 minutes and the SEM condenser lens excitation setting is set at 7 (beam current 2×10^{-11} A).

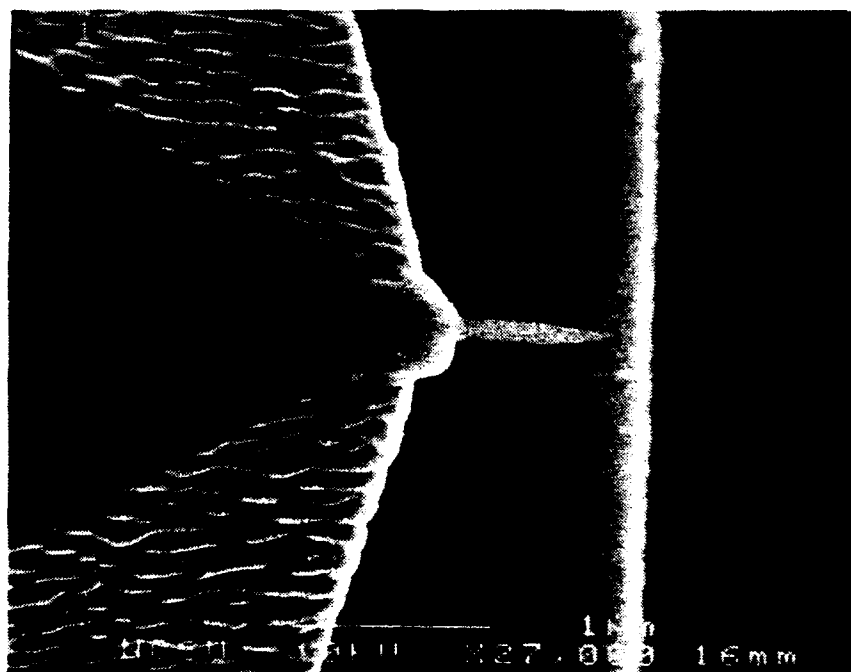


Figure 12: SEM micrograph of a microtip grown at an accelerating voltage of 15 kV. The growing conditions are identical to Figure 11.

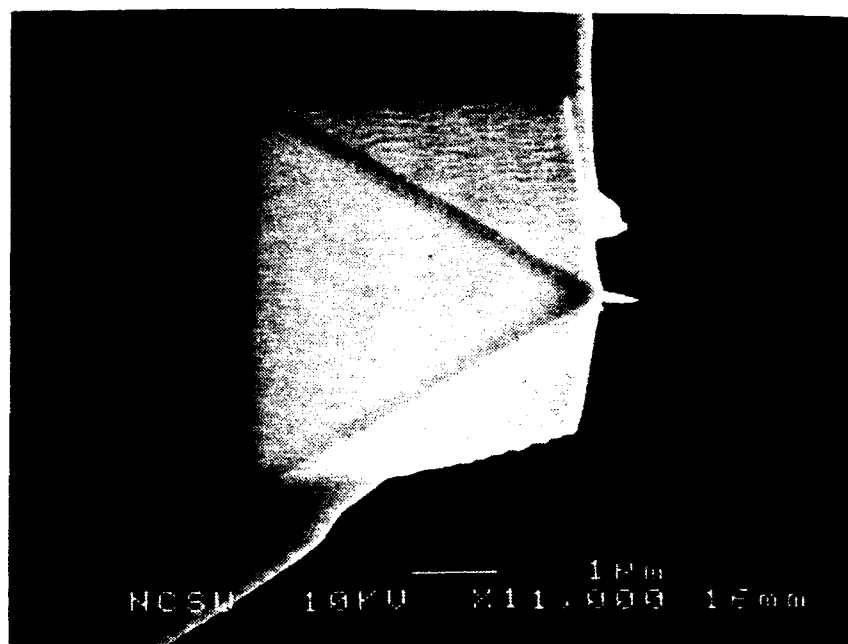


Figure 13: SEM micrograph of a microtip grown at $CL = 4$. The tip is grown for 3 minutes and the accelerating voltage is kept at 10 kV.

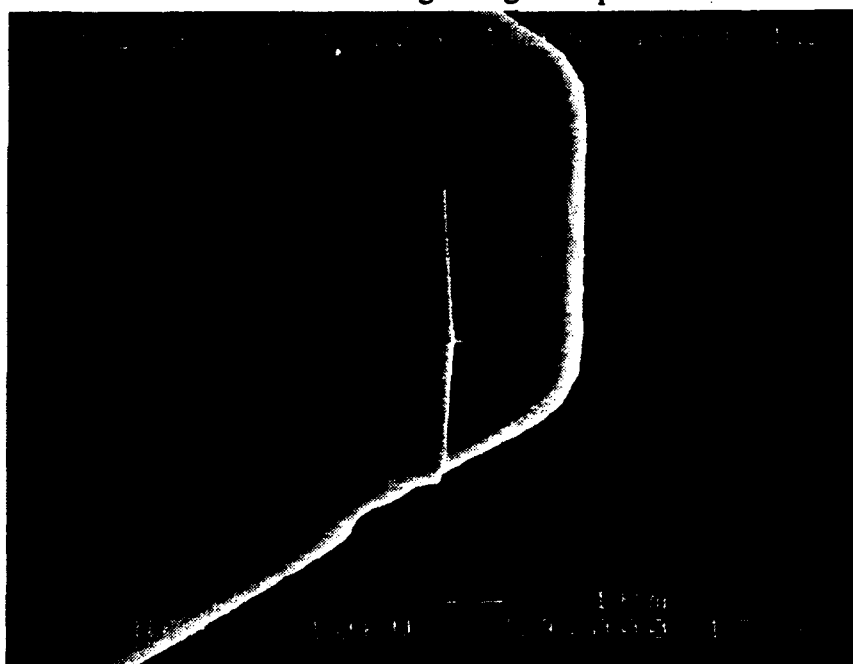


Figure 14: SEM micrograph of a microtip grown at $CL = 9$. The tip is grown for 1 minutes and the accelerating voltage is kept at 10 kV.

Among all the parameters that influence the tip geometry, the electron beam focusing is the most important. Since fine focusing is accomplished manually, the whole tip fabrication process has a relatively low throughput. Presently the parameters used for the tip fabrication are: 1) In metal coating: 70 Å thickness Au or Pt; 2) In FIB ion implantation: 20 kV Ga^+ beam, beam current 300 pA, 250 nm beam spot size and 1 minute implantation time; 3) In electron beam induced growth: 3

minutes beam irradiation time, 10 kV beam accelerating voltage and condenser lens excitation 7 (beam current 2×10^{-11} A). Although a sharper tip can be grown with a higher CL setting, the focusing process is more difficult since the final beam current is smaller, and thus the throughput is reduced.

3.3.4 AFM Imaging Results

To demonstrate improvement of commercial AFM tips with electron beam induced growth microtips, a series of AFM images of gold coated polymethylmethacrylate (PMMA) lithographic test patterns (gratings) have been obtained with a commercial (Nano-Scope II) AFM system by using AFM tips with and without the microtips. All AFM images have been obtained in air with minimum contact forces. Figure 15 shows a typical force curve as the microtip is scanned on a sample surface. It should be noted that the tip is operated at minimum force (zero deflection). A decrease in the force will cause the tip to lose contact with the sample surface or an increase in the force may damage the tip.

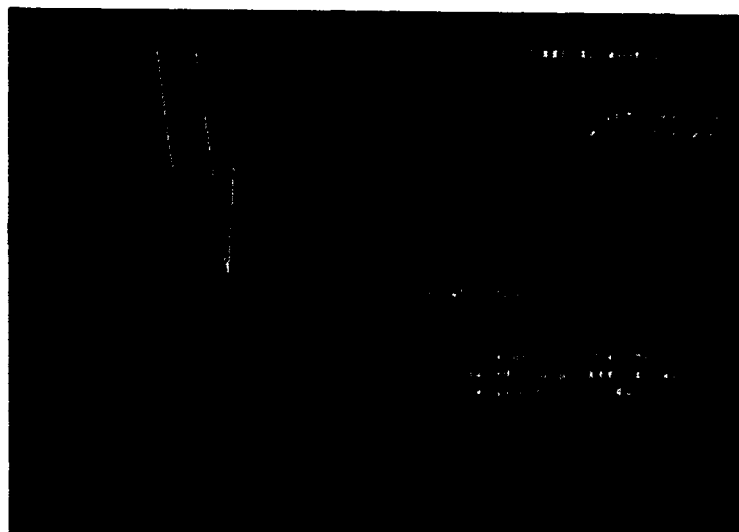


Figure 15: A typical force curve as the microtip is scanned on the sample surface.

For comparison, an AFM image of a grating has been taken with a pyramidal AFM tip, as shown in the three dimensional (3-D) perspective view of Figure 16, where pyramidal tip artifacts of the grating sidewall are apparent even for gratings with large periods. AFM images of gratings were obtained, as shown in the 3-D perspective view of Figure 17, where a relative straight sidewall has been mapped out by the AFM with the electron beam induced growth microtip fairly successfully. Furthermore, AFM imaging by microtips on two dimensional gold coated PMMA gratings and on gold coated surface are shown in Figure 18 and 19, respectively. It is apparent that the microtip is more successful in imaging large-scale, high-aspect-ratio pattern than conventional tips, however, the microtip has less significant effect on other type of samples.



Figure 16: AFM image of a gold coated PMMA grating by a commercial pyramidal tip,.



Figure 17: AFM image of a gold coated PMMA grating by a beam fabricated microtip.

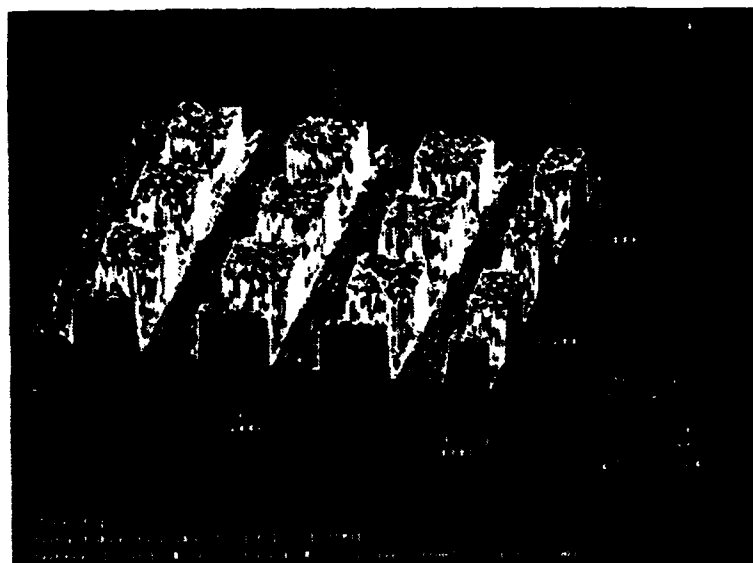


Figure 18: AFM image of a two dimensional gold-coated PMMA grating by a beam fabricated microtip.

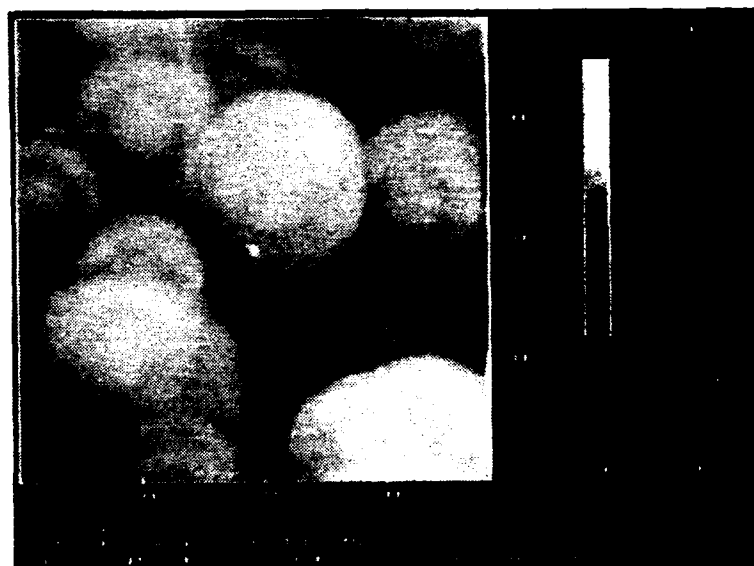


Figure 19: AFM images of gold-coated surface by a beam fabricated microtip.

3.4 CONTAMINATION TIP GROWTH

This growth method, that of simple electron beam induced contamination buildup, is similar to that described in the previous section and to all other electron beam induced growth methods, however, it does not involve the use of extra processing steps or a specially equipped SEM. This method could be the most promising since it requires only one step. However, for it to be a serious competitor, the tips must have dimensions similar to the other methods and not wear significantly over a reasonable amount of scanning time.

3.4.1 Tip Growth Parameters

Before growth the cantilevers must be metal coated on the side that the tip is to be grown on to avoid charging since beam energies are usually around 10 keV. The charging issue was discussed in more detail previously in this section.

Actual growth of tips is quite easy. Once the SEM is set up properly the scan generator for the beam is turned off, placing the microscope in the spot mode. The beam spot is placed at the point where a tip is desired and left for several minutes. This produces a cone of contamination, the shape of which depends upon the SEM parameters. Choosing parameters properly is important since different parameter values can produce a wide range of tip shapes. Microscope parameters include; beam energy or accelerating voltage, condenser lens excitation (controlling beam current and spot size), objective aperture size (controlling beam divergence and indirectly, current and spot size), and working distance. The parameters which are not controllable are the partial pressure of background hydrocarbons, and the amount of surface contamination. The last parameter is a bit misleading, since the amount of surface contamination can be controlled to some degree by sputter cleaning the pyramid or to begin with a clean pyramid. It is known that all of the cantilevers have at least a monolayer (and perhaps many more) of contaminants prior to insertion into the SEM.

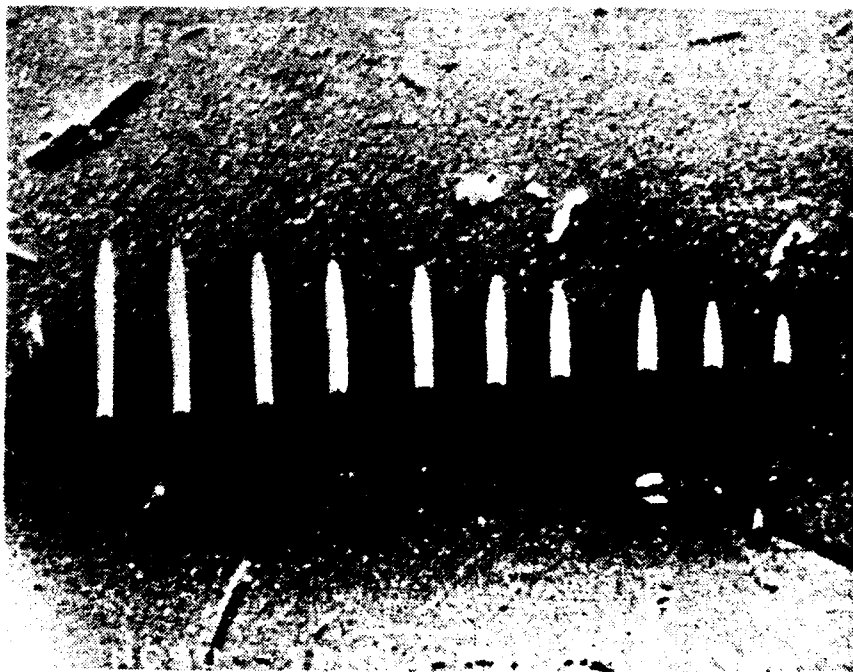


Figure 20: Micrograph showing contamination cones (or tips) grown for varying times. Starting with on the left, the growing time decreases from 5 minutes to 30 seconds in 30 second intervals.

Another user controlled parameter is the amount of time that the beam is left upon the chosen spot. The length of a tip is dependent upon how long the beam is left in a particular place. The time dependence is shown in Figure 20. Unfortunately, the stability limitations in the SEM render it impractical to grow straight tips for longer than four or five minutes. It is also not unreasonable to assume that the mechanical stability will decrease if the tip is too long. The current tested limit for tip length is 1.3 μm which is approximately 4 minutes of growth. The time factor is something that has to be adjusted each time that tips are grown, for the vacuum and thus the partial pressure of the background hydrocarbons is different. The only way of controlling this so far has been to grow a tip and see how long (or short) it is and adjust the time accordingly. This is not as difficult as it sounds, because the tip growing time is within one minute from one experiment to the next.

The working distance seems to be the least critical parameter, as long as it is kept reasonably short (<15mm) and the accelerating voltages are kept reasonably high (>5kV), with no observable differences noted between tips grown at working distances between eight and fifteen millimeters with accelerating voltages above 5kV. The working distance becomes significant as the accelerating voltage decreases, since any external fields affect the low energy beam to a much greater extent due to the increased travel gap between the final lens and the specimen.

As long as the objective aperture is relatively small (30-50 microns) it does not seem to play too great a role since the spot size is much smaller than the size of the grown tip. As of this writing the objective aperture parameter has been the least thoroughly investigated with only the smallest objective apertures having been used to grow tips that have been used in the AFM. It can be seen from Figure 21 that the objective aperture does play an important role in tip size and shape with the largest two apertures (70, 110 microns) producing tips that are clearly not desirable. The tips that have been used to take images in the AFM have all been similar to the tip (growing conditions identical) on the far right. The sharpest tips, the one on the left, was grown with objective aperture 4 (30 microns) and condenser lens 8. No tips that have been grown similar to this tip have been used successfully to image with. Those that have been tried have broken in the first scan, however, proper handling procedures had not been established when they were used and it is unknown whether they will work.

The shape of the tip is highly dependent on the accelerating voltage. At accelerating voltages of 1 to 2 kV, the tip is large and not very sharp. At high accelerating voltages of 20 kV, the tip becomes more cone-like and the aspect ratio decreases. The best tips have been grown between 5 and 15 kV with the majority grown at 10 kV. Unless otherwise noted all of the tips in this section have been grown at 10 kV.

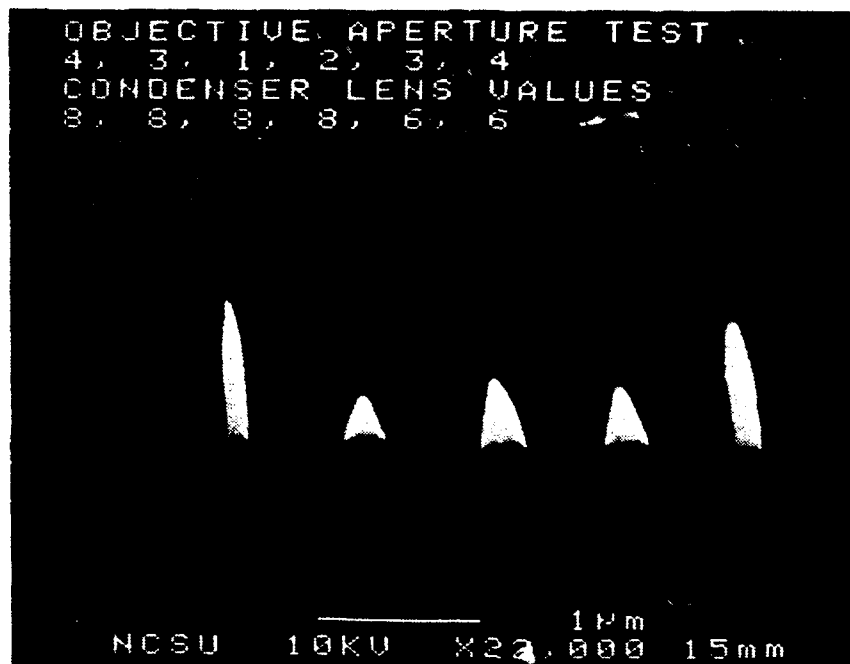


Figure 21: Experiment to see the effects of changing the objective aperture. The apertures sizes are 1 = 110 microns, 2 = 70 microns, 3 = 50 microns, 4 = 30 microns. The condenser lens had to be increased in the middle of the experiment because no tips formed at a condenser lens value of 6 (on a scale of 1-10) for the two smallest apertures. The growing time (3.5 minutes), accelerating voltage (10kV), and working distance (15mm) are all held constant for the experiment

The excitation of the condenser lens seems to define the diameter of the tip shaft and perhaps the end point as well. This is not surprising since the condenser lens defines the diameter of the electron beam. The effect of the condenser lens excitation can be seen in Figure 22. At low condenser lens settings the tips do not grow well and are too large too use. At very high condenser lens settings the tips have extremely small diameters, however, they do not appear to be mechanically stable enough for use (none have been imaged with successfully thus far). Favorable growth is found at condenser lens settings about midway between either extreme. These mid-range (5-7 on a scale of 1-10) condenser lens excitation values produce straight sharp tips that are mechanically stable for general use. Another point to consider is the signal to noise ratio as the condenser lens is excited. The ratio decreases as the condenser lens is excited since the current decreases at higher condenser lens settings. The result is that the image is difficult to see and thus properly focus and stigmatize.

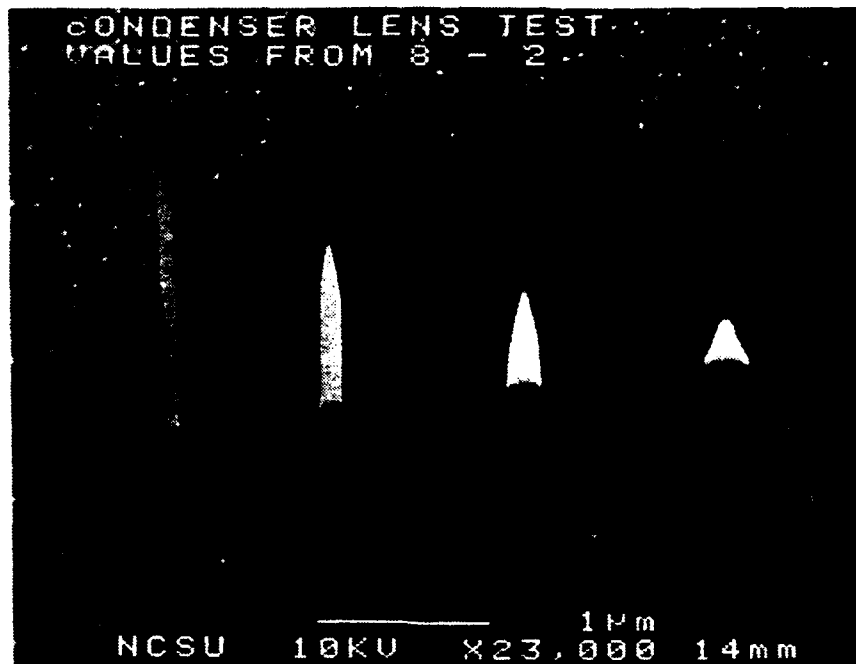


Figure 22: Condenser lens parameter experiment. The condenser lens varies from 8 to 2 from left to right. The other parameters were held constant at the following values; accelerating voltage = 10kV, working distance = 15mm, objective aperture = 4 (30 microns), and the growth time = 3.5 minutes



Figure 23: Microtip that was grown with improper stigmation. Note that it is elliptical instead of circular.

Utmost care needs to be taken in correctly focusing and stigmating the SEM for the shape of the tip is dependent upon how well this is done. If the SEM is not properly focused the tip is larger than it would normally be. It may also have an odd, unpredictable shape and may not be as mechanically strong as a tip grown with proper focus. Stigmation is also very important since an improperly stigmated beam is not circular and nor are any tips grown with such a beam. A tip grown with proper stigmation will be uniformly circular when viewed from above and should give the best images in the AFM. Figure 23 shows a tip grown with improper stigmation.

3.4.2 Tip Growth Methods

A contamination cone (tip) forms when the electron beam is left in one place for a period of time. The cones grow in the direction of the beam since the beam is providing the energy necessary to polymerize the hydrocarbon background vapor. Since these tips grow up in the direction of the beam, they are referred to as "up" or normal tips. All of the contamination tips shown thus far have been grown by this method. The geometry for growing tips in this manner is quite simple. The pyramids are placed at 80° from normal, or 10° from the horizontal as in Figure 24. The 10° offset is to correct for the 10° angle at which the tips are held in the AFM. This produces tips which will image properly. The vast majority of the tips grown thus far are grown in this fashion for the simple reason of easy geometry. An SEM micrograph of the proper position is shown in Figure 25.

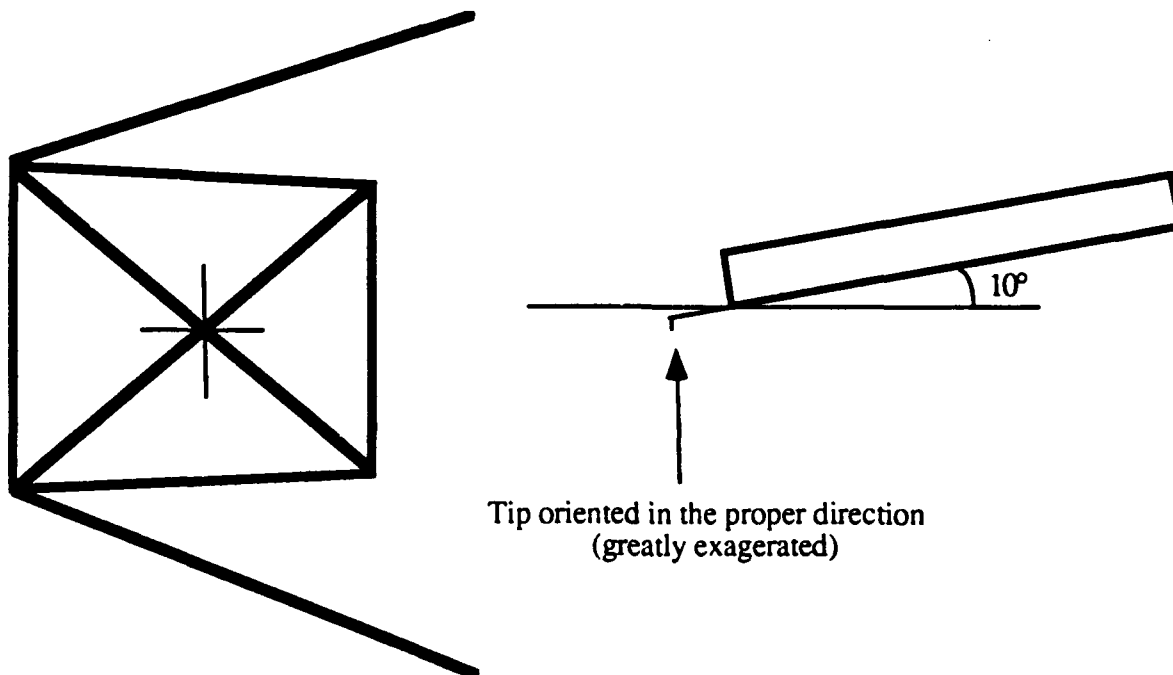


Figure 24: Geometry of normal contamination tip growth. The tip is grown up (out of the paper) at the point where the cross-hairs are. The smaller figure shows that this produces the proper orientation for the tip.

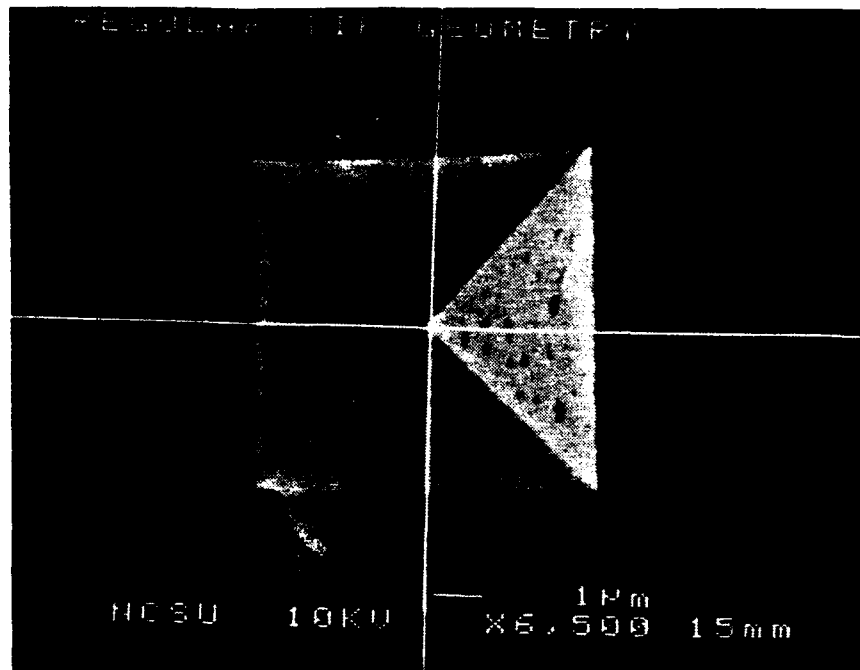


Figure 25: Micrograph Illustrating geometry in Figure 4.2.1

There is another method of growing tips that is referred to as "side growth.". This method differs from that above in that the beam is moved in the direction of the tip instead of being left stationary while the tip grows towards it. In other words, a side view of the pyramid is found and the beam is placed at the desired point on the tip of the pyramid and then the beam is moved outward to form the tip. This presents a new geometrical problem since the beam is not only growing in the direction of beam motion, but also in the direction of the beam. To better understand this Figures 26, 27, and 28 below. The rate at which the tip grows in the direction of the beam seems to be constant, however, the rate at which the SEM operator moves the spot may not be. This introduces another difficulty in estimating the amount of offset needed to correct for the growth in the direction of the electron beam. This method is promising in spite of the difficulties since the tips can be grown in a variety of shapes for imaging surfaces that would normally be impossible with a standard tip.

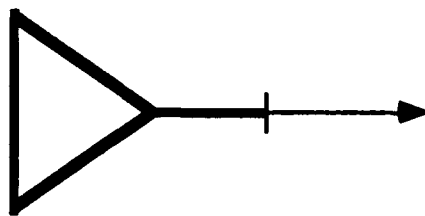


Figure 26: Geometry of "side" grown tips illustrating the difficulties encountered. As the beam is moved in the direction of the arrow, a tip is grown (the cross-hairs represent the point at which the beam is located). However, as the tip grows in the direction of the arrow it is also growing up (out of the paper), resulting in a tip similar to the one in Figure 27.

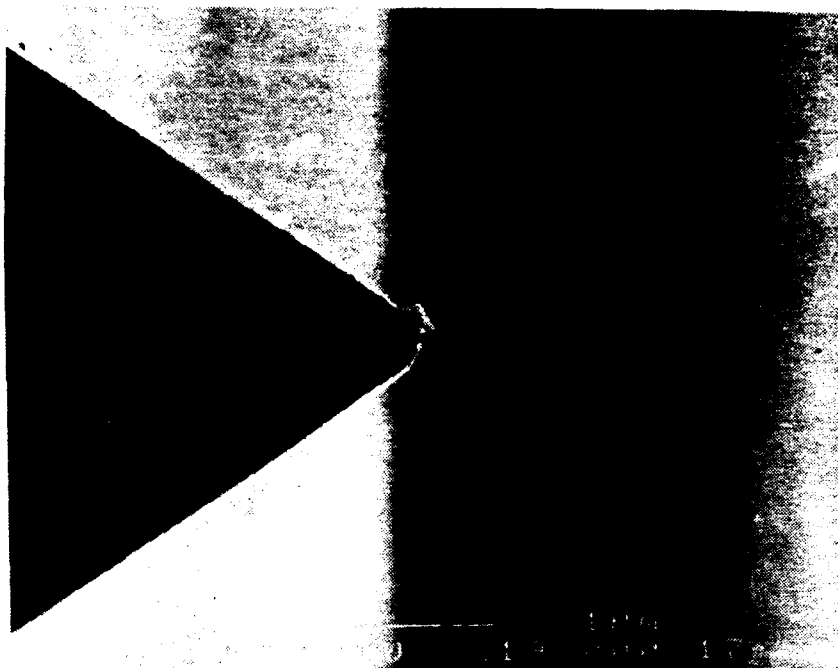


Figure 27: This tip looks very good from this angle. When viewed from above (in Figure 28), however, it is obvious that it is not so good.

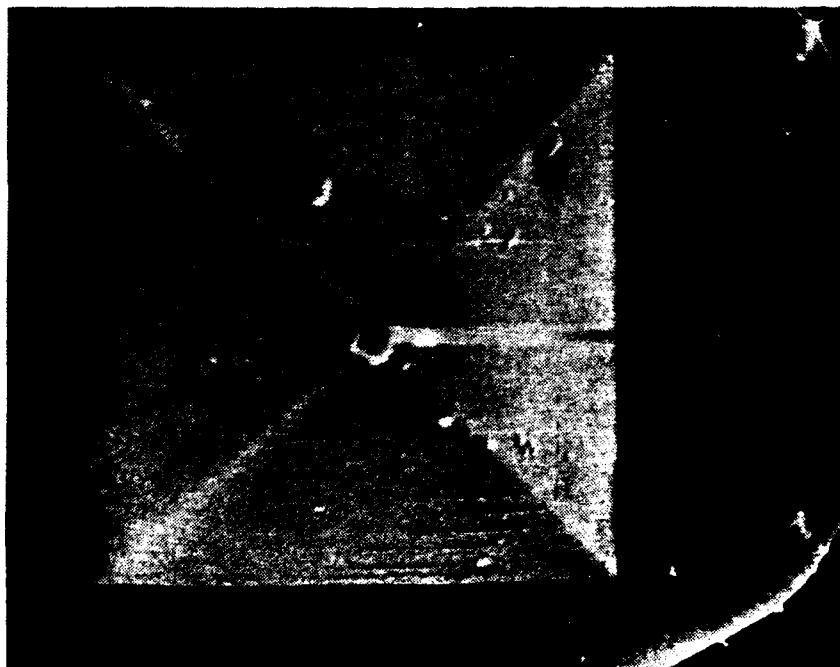


Figure 28: The same tip as in Figure 27 when viewed from above. It is all too easy to become misled by what appears at first glance to be a good tip

3.4.3 Contamination Tip Imaging

Tip Handling . e most difficult thing to learn about imaging with grown tips is proper handling of the micro-tips. Standard tips are very tough and can be dropped with no ill effect, unless of course the tip lands directly on the cantilever. Grown tips are much more sensitive to dropping and may break very easily. Other than dropping a tip, separating the individual cantilevers from a strip of cantilevers seems to be the most important factor. The standard cantilevers are formed on a wafer that has many strips of cantilevers. The strips are scribed so that the individual tips may be broken apart down preset lines such that the individual cantilevers fit into the AFM tip holder. How gently or violently the individual tips are broken off seems to be of great importance since many tips are believed to have been damaged by improperly separating the individual cantilevers from the strip. Some of the cantilevers that were broken improperly were checked in a SEM to see if the grown tips still existed. In most cases the tips seemed to be intact but failed immediately when used in the AFM. This effect remains unexplained. Once the proper procedure for separating individual cantilevers was found imaging was quite easy. Once the tip has "engaged" the sample the force that is used has not been observed to be too important. While this should not suggest using high tip to sample forces this does suggest that the average AFM user could successfully use these tips provided that the cantilever separation was done properly.

Imaging Data The standard test pattern for characterizing the performance of an individual tip is a lithography pattern consisting of "towers" and "trenches" of various widths. The pattern is undercut so that no excess sidewalls exist. This pattern is ideal for characterizing the large scale (>100nm) imaging capabilities of the tips. Figure 29 shows the standard lithography pattern used for tip testing imaged with a standard tip. Figure 30 shows a similar area imaged with the grown contamination tips (the area of the "tower" is larger, however, the slope of the sidewalls and the sidewall height is the same).

Due to both hardware and software limitations, right angles are too steep to measure with a Nanoscope II AFM. In this case, 90° angles would represent two data points in the same space which is not allowed. This also prevents the measurement of undercut sidewall lithography patterns. Even though it is possible to grow tips that are capable of imaging undercut sidewalls the hardware and software are not at this time commercially available to do so. The limit for sidewalls with a low scan rate (<2Hz) and a scan size of <10 microns over a step of 1 micron is approximately 88°.

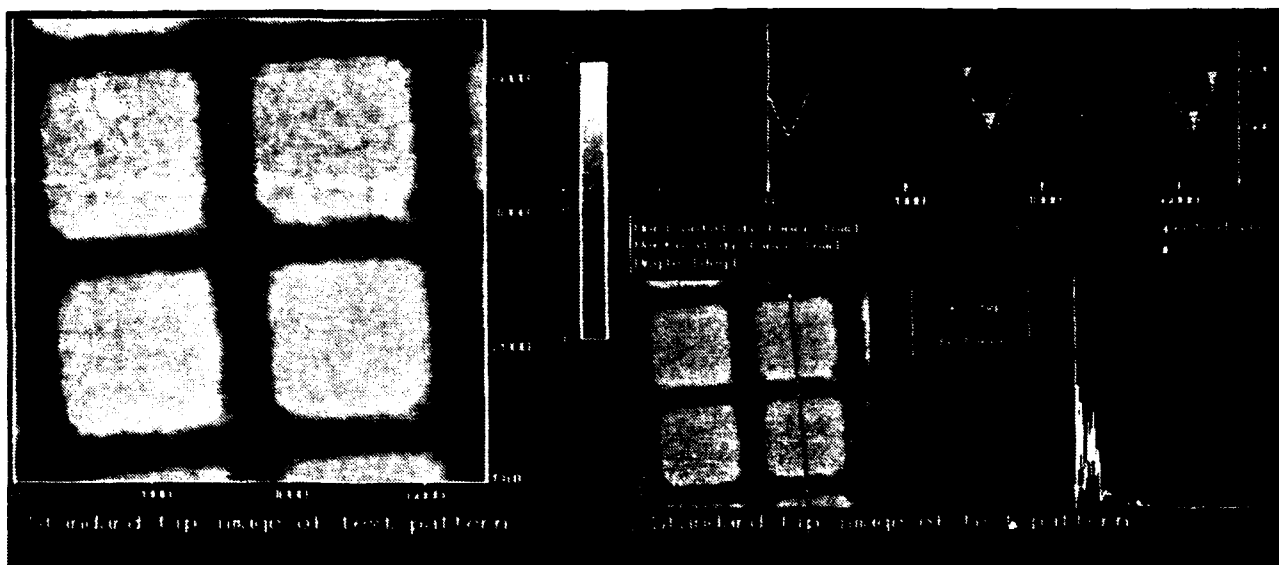


Figure 29: Undercut lithography pattern imaged with a standard AFM tip. Note how the sidewalls slope off.

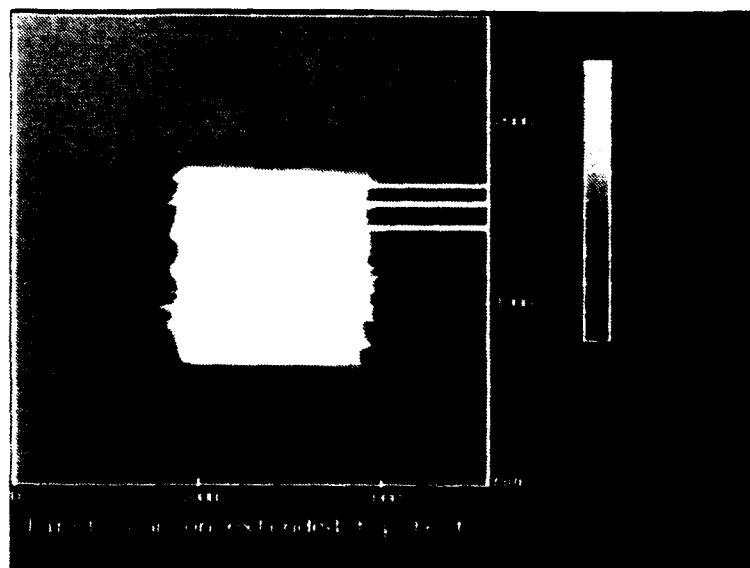


Figure 30: Undercut lithography sample imaged with a grown contamination microtip. Note how sharp the sidewalls appear.

Longevity Data Contamination microtips have been used to image for up to 48 hours on end with little or no wear. One of the contamination microtips has been tested for longevity and was left to image for an entire weekend. At the end of this period no detectable wear can be seen (see Figure 39). That little or no wear took place is rather surprising since the piezo scanning tubes on

the Nanoscope II tend to drift over time usually resulting in an increase in scanning force. Several times over the 48 hour period the force was found to be on the order of forces used in fibril manipulation experiments. It was thought prior to this experiment that forces of that magnitude would break the tip, however, the high forces associated with the Z-piezo drift did not appear to adversely affect the tip. These points are important since useful tips must be durable and have a relatively long lifetime in order to be economically feasible for the average lab. Unfortunately this tip had some irregularities, most noticeably a bump on one side. The effects of this irregularity on the images can be seen in the figures below illustrating how important a uniform, symmetric tip is if one is to properly interpret the images with no prior information about the particular tip used.

Figure 31 shows a top-view of the lithography pattern during the beginning of the extended test. Figures 32 and 33 show the angles that are found in the "section" mode in the fast and slow scan directions respectively. Note that the steepest angle in Figure 33 approaches the theoretical limit of the Nanoscope II software. Figure 34 is a linescan image of the data in Figure 31. Figure 35 shows a top view of the last scan, after 48 hours of continuous imaging. Figures 36 and 37 show sectioned views in the fast and slow scanning directions respectively. Figure 38 shows a linescan of the data in Figure 35. Figure 39 is a SEM micrograph of the tip used to image for 48 hours. Note the irregularities in the tip that show up in the AFM images.

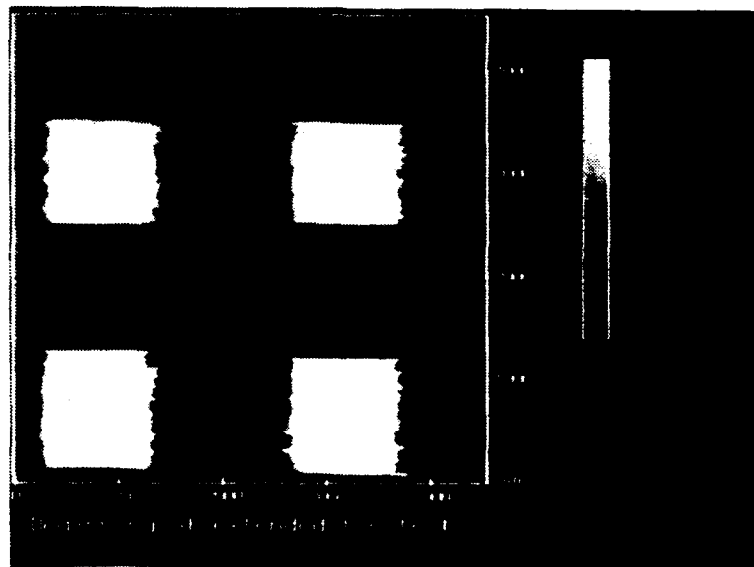


Figure 31: Image showing the topview of a scan of the undercut lithography sample at the beginning of an extended (48 hour)-test.

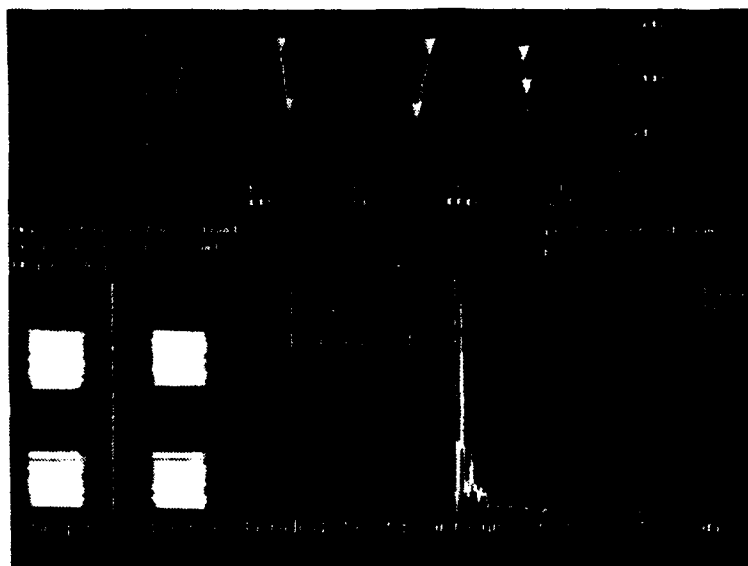


Figure 32: "Section" showing the angles obtained in the fast scan direction in Figure 31.

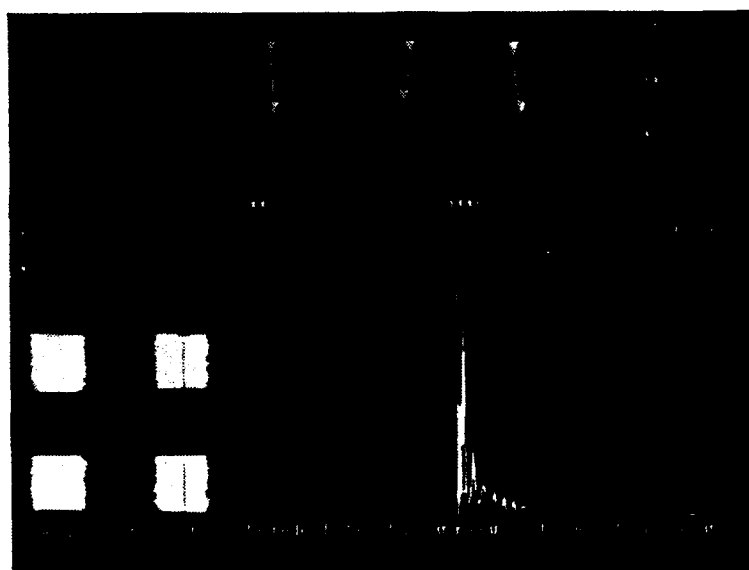


Figure 33: Figure showing the angles obtained in the slow scan direction of Figure 31.

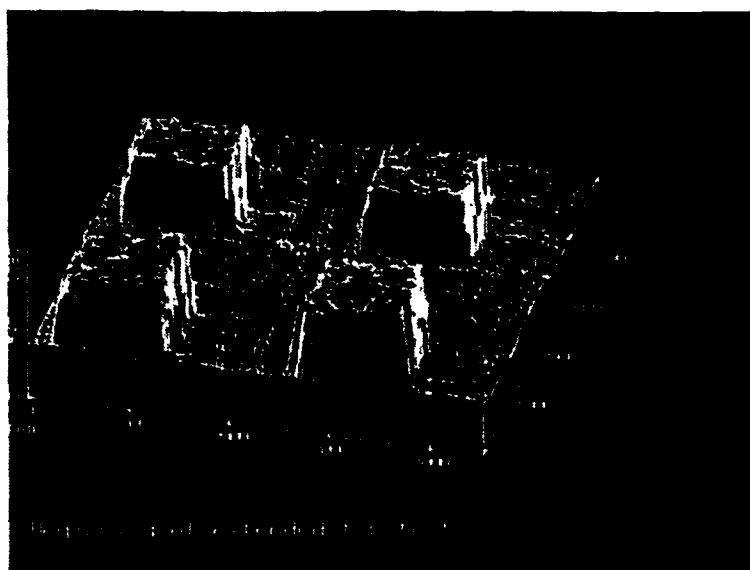


Figure 34: Linescan of Figure 31.

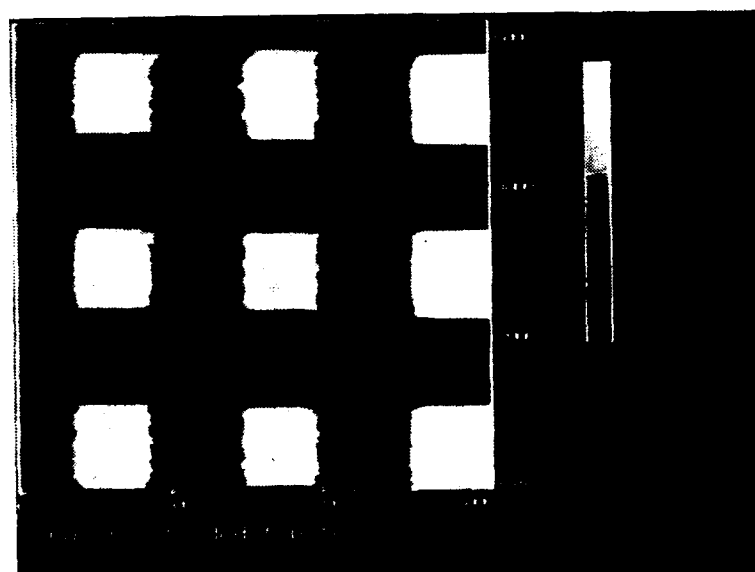


Figure 35: Topview of the last scan, after 48 hours of imaging with the same microtip.

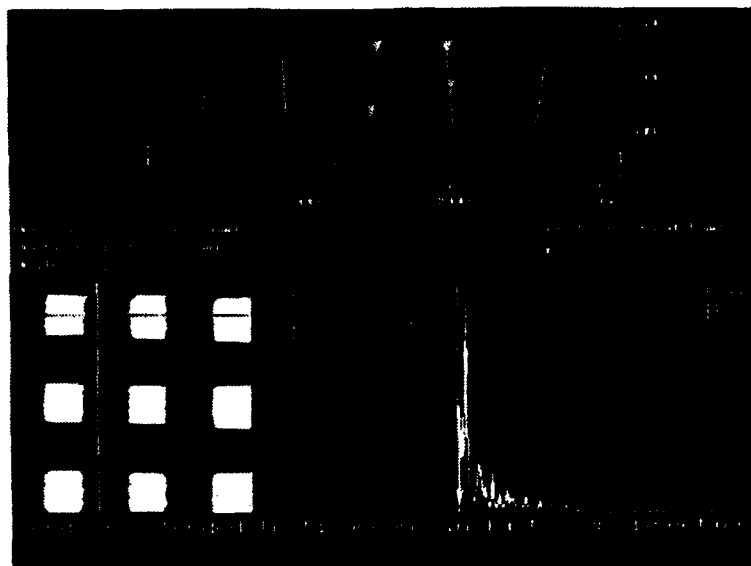


Figure 36: Section showing the angle obtained in the fast scan direction of Figure 35.

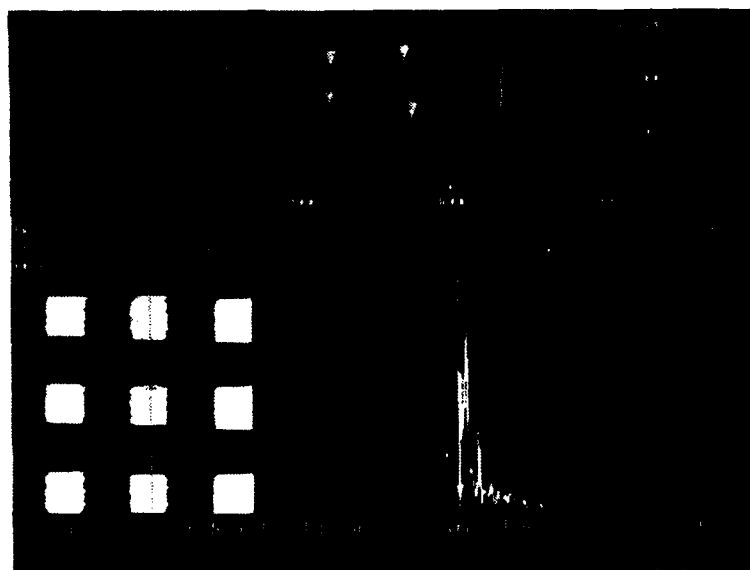


Figure 37: Section showing the angles obtained in the slow scan direction of Figure 35.

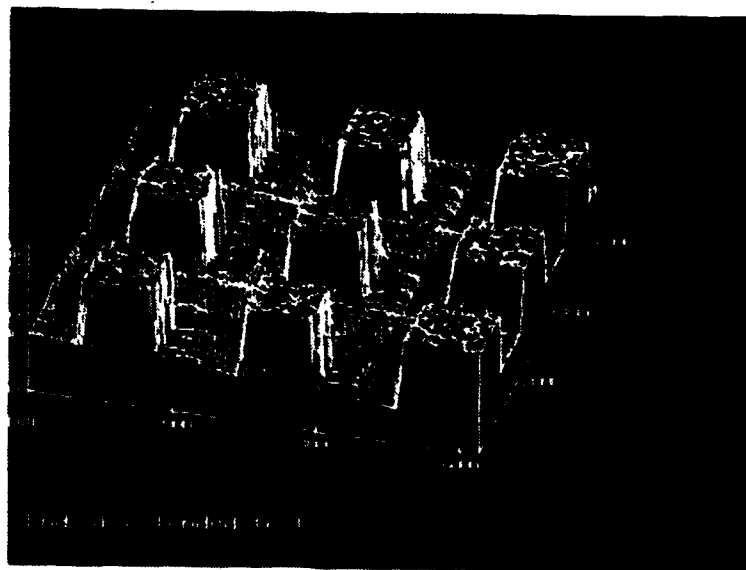


Figure 38: Linescan of Figure 35.

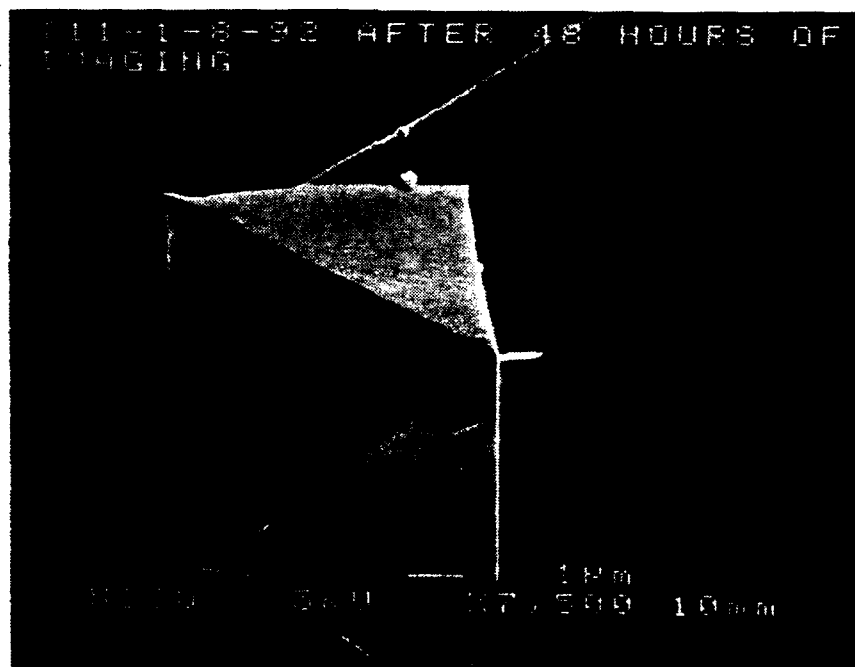


Figure 39: The contamination tip that was used to image the lithography test pattern for 48 hours. Little or no wear is seen. Note how the irregularities of the tip can be seen in the images taken with the tip.

3.5 AFM IMAGING AT ATOMIC SCALE USING MICROTIPS

Microtips have been successfully used for imaging large-scale, high-aspect-ratio samples. However, it appeared to be impossible to use these tips for atomic resolution imaging. AFM imaging of mica has been attempted with over 13 tips with the tip length of 0.5 μm to 1 μm . In the

process, the microtips were operated at high forces (both repulsive and adhesive), but, atomic scale images could not be achieved. It may be that tips of these lengths are too flexible because of their length. Instead of sticking to the sample surface to sense the force between itself and the atom, it may be deflecting away from the scan direction. Therefore, to image at atomic scale, the tip should be very short, at least less than 100 nm [18].

Since the tip length can be controlled by varying the electron beam irradiation time short tips were made to check the deflection theory. However, it is not easy to control the tip length within 10 nm. By carefully adjusting all the parameters, a 20 nm long tip was made by stationing the electron beam for 2 seconds on the top of the pyramid. As shown in Figure 40, the end radius of tip is about 15 nm. Initial attempts at AFM imaging of mica sample using this tip is promising. SEM observation proved that the tip was still good after the AFM imaging, as shown in Figure 41. A comparison of AFM image without any data processing using 0.5 μm long tip and 20 nm tip are shown in Figures 42 and 43. In the image made by 0.5 μm long tip only noise can be seen, while in the one made by the 20 nm tip a clear atomic picture is shown. A clearer atomic scale picture (after filtering) imaged by the 20 nm tip is shown in Figure 44.

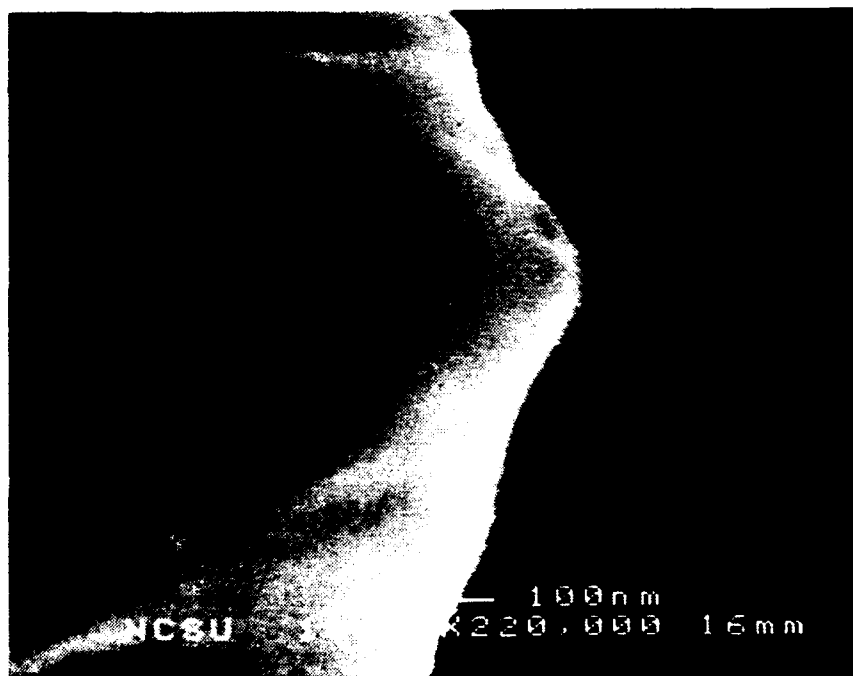


Figure 40: A 20 nm long microtip, made by dwelling the electron beam on top of the pyramid for 2 seconds, before AFM imaging. The end radius of the tip is 15 nm.

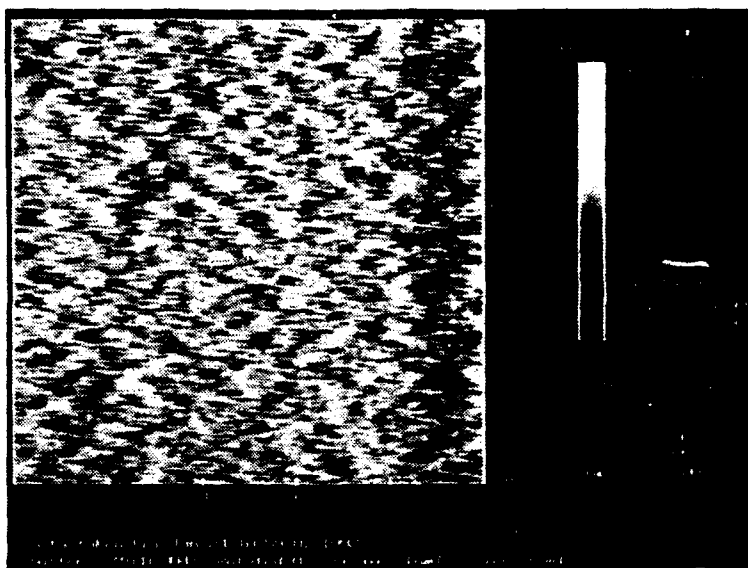


Figure 43: Atomic resolution AFM images of mica sample taken with a 20 nm long tip. The image is raw data without any data processing. The image by 20 nm tip clearly shows atomic resolution.

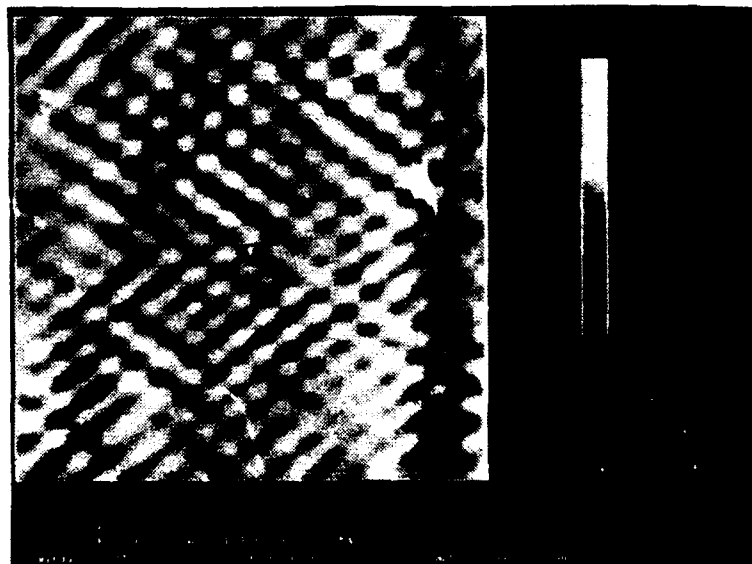


Figure 44: AFM image of mica sample (original data: Figure 40) after two-dimensional FFT filtering.

3.6 CONCLUSION AND FUTURE WORK

Several methods of improving commercial pyramidal Si_3N_4 AFM tips by FIB cutting and electron beam induced growing microtips on the top of either the Ga^+ implanted or metal coated pyramidal AFM tips have been developed. The growth processes are repeatable and produce tips which can be used to nearly eliminate the tip artifacts associated with standard Atomic Force Microscopy tips in the imaging of high-aspect-ratio samples. These tips have been used successfully to image a variety of materials with a reasonable lifetime. The geometry of the tip can be controlled by varying electron beam irradiation time, beam energy, and current density. Initial attempts of atomic resolution imaging using grown tips has been demonstrated.

The basic technology of the growth of microtip has been transferred to Material Analysis Service, Inc. (MAS). The microtips has also been sent to selected PEC affiliates who have tested the reliability and usefulness of the tips on their samples. Some users have had no success with microtips, while others easily reproduced our results. Shipping and handling of the microtips is still a critical point that must be addressed. The affiliates who have received tips have also found that proper handling is of utmost importance. Microtips were also fabricated on Digital Instrument's stand-alone AFM tips.

Preliminary energy dispersive (EDS) X-ray spectra have been taken during tip growth and the results show carbon contamination builds up in both types of tips. Further investigation of chemical composition, growth mechanism and stiffness of the grown microtips is underway. Direct topographic measurement of non-conducting microfabricated lithographic patterns could become one of the most promising applications for using AFM with the microtips. Efforts on imaging at atomic scale with grown microtips will continue.

Growth of various shape of tips by computer control of the electron beam direct writing will also be investigated. Direct growth of contamination tips will be thoroughly studied since it reduces the three-step process by one step. The current microtip fabrication procedure suffers from low throughput because it is a sequential process, therefore, attempts at increasing the throughput of this fabrication procedure will be initiated.

References

- [1] Albrecht, T.R., Quate, C.F. "Atomic resolution with the atomic force microscope on conductors and nonconductors." *J. Vac. Sci. Technol.* **A6**, (1988) 271.
- [2] Albrecht, T.R., Carver, T.E., Quate, C.F. "Microfabrication of cantilever styli for the atomic force microscope." *J. Vac. Sci. Technol.* **A8**, (1990) 3386.
- [3] H. Ximen and P. Russell, "Microfabrication of AFM tips using focused ion and electron beam techniques", *Ultramicroscopy*, In press 1992.
- [4] H. Ximen and P. Russell, "Atomic Force Microscopy Using Beam Fabricated Microtips", *Precision Engineering Interim Report*, 1991.
- [5] Hillier, J. "On the Investigation of Specimen Contamination in the Electron Microscope." *J. Appl. Phys.* **19** (1948) 226.
- [6] Ennos, A.E. "The origin of specimen contamination in the electron microscope." *Brit. J. Appl. Phys.* **4** (1953) 101.
- [7] Ennos, A.E. "The sources of electron-induced contamination in kinetic vacuum systems." *Brit. J. Appl. Phys.* **5** (1954) 27.
- [8] Fourie, J.T. "Contamination Phenomena in Cryopumped TEM and Ultra-high Vacuum Field-Emission STEM Systems." *Scan. Electron Microsc.* 1976. (1976) 53.
- [9] Fourie, J.T. "High contamination rates from strongly adsorbed hydrocarbon molecules and a suggested solution." *Optik.* **52** (1978) 91.
- [10] Linders, J. and Niedrig, H. "Development of microcones induced by contamination lithography." *Nuclear Inst. and Meth. Phys. Res.* **B13** (1986) 309.
- [11] Kreuzer, P. "Formation and examination of self supporting contamination filaments." *Optik* **78** (1988) 158.
- [12] Fuji, T. et al. "Micropattern measurement with an atomic force microscope." *J. Vac. Sci. Technology.* **B9** (1991) 666.

- [13] Lee, K.L., Abraham, S.F., Secord, F., Landstein, L. "Submicron Si trench profiling with an electron-beam fabricated atomic force microscope tip." *J. Vac. Sci. Technol.* **B9**, (1991) 3562.
- [14] Nyyssonen, D., Landstein, L., Coombs, E. "Two-dimensional atomic force microprobe trench metrology system." *J. Vac. Sci. Technol.* **B9** (1991) 3612.
- [15] H. Ximen and P. Russell, "Focused ion beam micromachining", *Precision Engineering Annual Report*, 1990.
- [16] Prater and Hansma, "The Scanning Ion-Conductance, Microscope", *Science*, Vol. 243, pp. 641-643.
- [17] J.G. Pellerin, G.M. Shedd, D.P. Griffis, and P.E. Russell, "Characterization of focused ion beam micromachined features", *J. Vac. Sci. Technol.* **B7** (1989) 1810.
- [18] D. Grigg, AT&T Bell Lab, private communication.

4 PROCESS AUTOMATION FOR PRODUCTION OF CONTROLLED GEOMETRY STM TIPS

James L. Robb IV

Undergraduate Student

Mechanical & Aerospace Engineering

Dieter P. Griffis

Research Associate

Analytical Instrumentation Facility

John M. Mackenzie, Jr.

Associate Professor

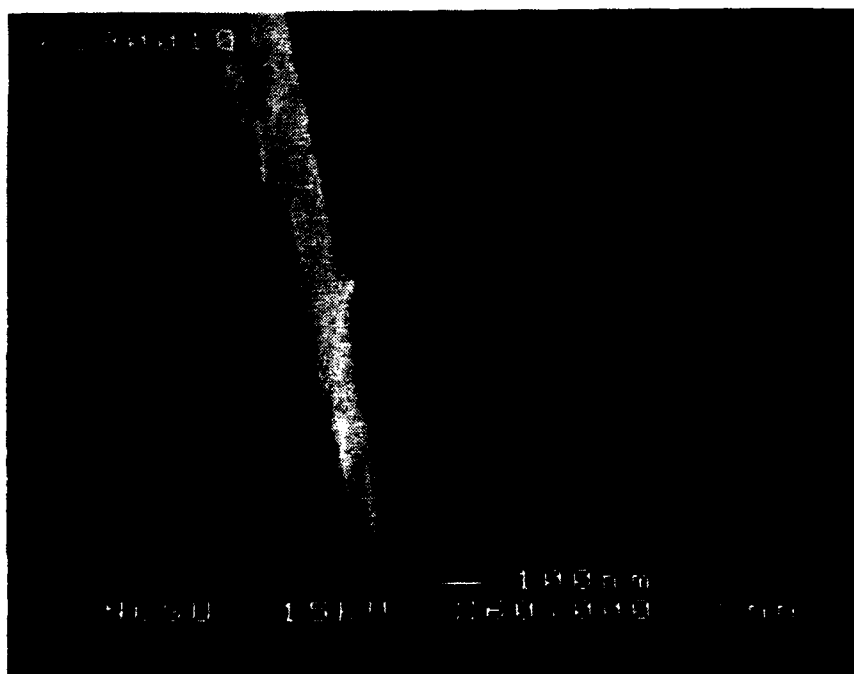
Department of Microbiology

Phillip E. Russell

Professor

Department of Materials Science and Engineering

An integrated system for automated production of controlled geometry STM tips is under development. An automated bulk etching procedure has been successfully implemented which produces tips suitable for polishing. The polishing procedure has produced tips having a radii less than 50nm .



4.1 INTRODUCTION

The sharp, controlled geometry Pt/Ir STM tips produced using the patented process developed by Musselman and Russell [1,2] provide dramatic improvement in the capability of the STM to measure high aspect ratio surface features. These tips, now manually produced using the above process and marketed by Materials Analytical Services, Inc. (MAS), are in great demand in the U.S. and internationally. The manual process requires two separate apparatuses, one for bulk etching and one for polishing. In addition, manual production of these tips, which requires repeated critical judgements to be made by the operator, has proven to be extremely tedious. These factors effectively limit the number of tips that can be produced and impact unfavorably on quality control. To increase production capability and improve quality control, a simplified (one system for both bulk etching and polishing) system is under development and critical aspects of tip production are being automated in order to limit human intervention to mixing of solutions and removing tips from the apparatus for rinsing and placement into protective containers. Enhancement support has been provided by MAS for this automation process to improve the productivity of their manufacturing.

4.2 AUTOMATION SYSTEM

4.2.1 Tip Etching Apparatus

A schematic of the tip etching automation system design is shown in Figure 1. The primary component of the apparatus is a computer controlled micropositioner with $0.1\mu\text{m}$ positioning resolution connected to a linear stage to provide the vertical motion of the tip. The base of the stage is mounted to the inside wall of a plexiglass tube. A bracket bored to provide a feed for the Pt/Ir wire is attached to the stage. The tube containing the positioner is then mounted to another concentrically-aligned plexiglass tube with an equal OD. By having a larger wall thickness for the bottom tube, a plate can be placed between the two tubes thus creating two chambers and allowing isolation of the positioning equipment from the corrosive etchant.

The lower tube chamber contains the etchant necessary for tip fabrication. The etchant container and finished tips are removed through two windows in the bottom tube. Both the upper and lower chambers of the tube system are purged with N_2 . The upper chamber is purged to

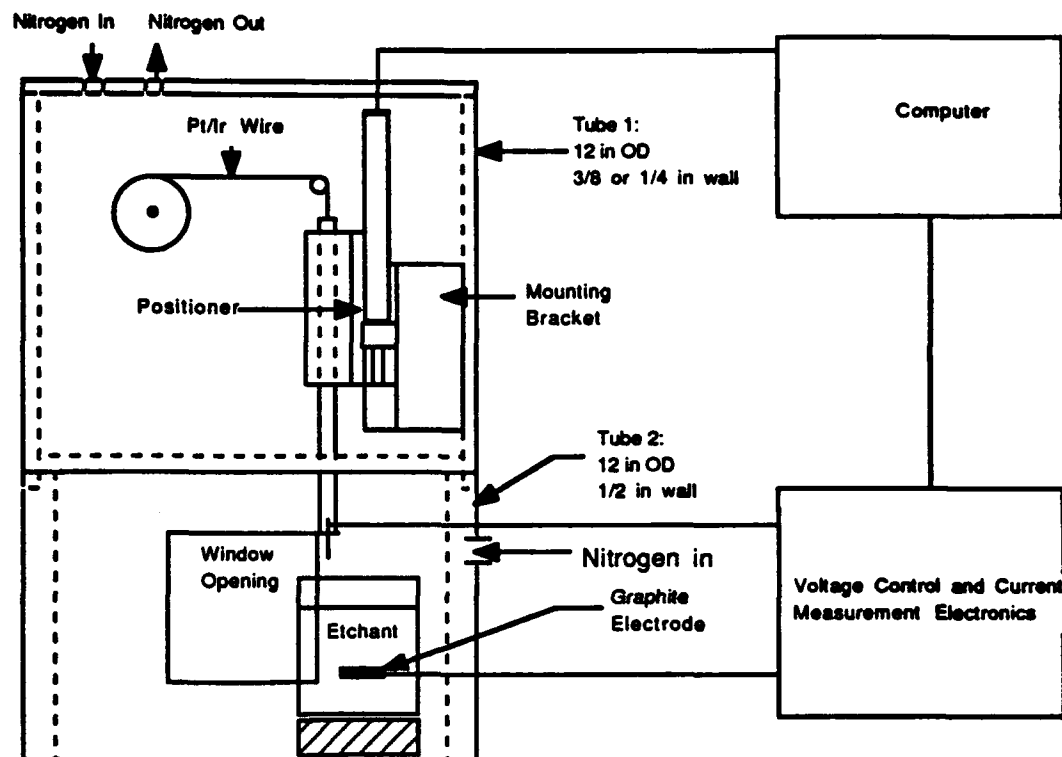


Figure 1: Schematic of Pt/Ir Tip Etching Apparatus

protect the positioning mechanism from the corrosive CaCl_2/HCl solution used as the etchant. The lower chamber is purged to prevent atmospheric CO_2 from being absorbed by the CaCl_2/HCl solution to prevent the formation of a carbon film on the tips during etching which interferes with tunneling making the tips unusable.

4.2.2 Control Software

Software for the automated control of the above system has been written and testing of the combined mechanical positioning, voltage switching and current measurement system has begun. The software written is written in BASIC for ease of modification, and subroutine calls to C provide low level positioner control. The software integrates tip positioning, voltage switching and current detection into a control package which will allow unattended etching of tips throughout the etching process. At this stage of development, human intervention will be needed for loading and unloading of the tips. The algorithm used for tip etching described below.

4.3 AUTOMATED TIP ETCHING

The effort to determine the process parameters required to reproducibly fabricate STM tips having tip radii less than 50nm is divided into two separate processes. First the 0.254mm O.D. Pt/Ir wire must be bulk etched into the tip shape required for polishing. The result of the bulk etch is then polished to the required sharpness. The solution formulation of the CaCl_2/HCl solution used in this process is unchanged from that described by Musselman [2]. Both the bulk and polish etching processes described below are carried out sequentially in the same solution with the Pt/Ir wire providing one electrode and a carbon rod the other. The entire process, after initiation of the bulk etch by the operator, is under computer control until the polishing process is completed.

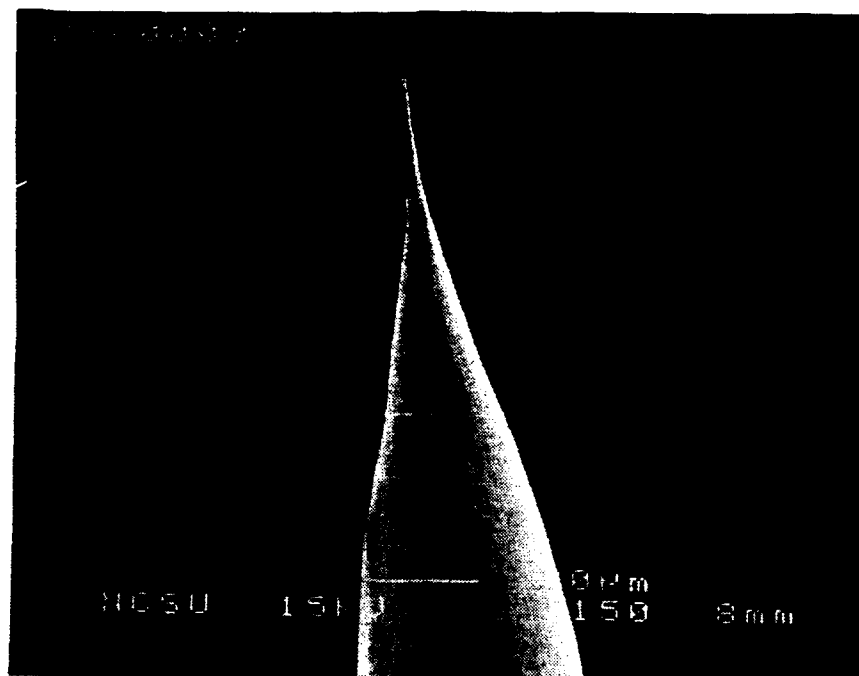
4.3.1 Automated Bulk Etching Process Sequence

The results of a typical bulk etch is presented in Figure 2. Upon initiation of the process by the operator, the 27.5VAC bulk etching voltage is switched on by the computer and the unetched Pt/Ir wire is moved toward the surface of the solution by the linear DC motor. Upon sensing current flow due to wire contact with the surface, the voltage is shut off and the wire is inserted 1.5mm below the surface of the solution. For the automated process both the voltage and the depth of immersion was found to be critical to provide the proper tip geometry for subsequent automated polishing (discussed below). The etching voltage is then switched on and the bulk etch continues to completion. Completion is detected by sensing the large current drop when the wire is etched into 2 pieces. The wire is progressively thinned near the surface until it is severed by melting due to high current flow. The result of this severing due to melting can be seen in Figure 2(b). The progressive thinning of the wire near the surface provides the approximate geometry required for STM measurement of high aspect surfaces requiring only sharpening by polishing. This process is sufficiently well characterized that an analog system for bulk etching has been designed and implemented and two such systems are currently operated by MAS to bulk etch all tips prior to manual polishing.

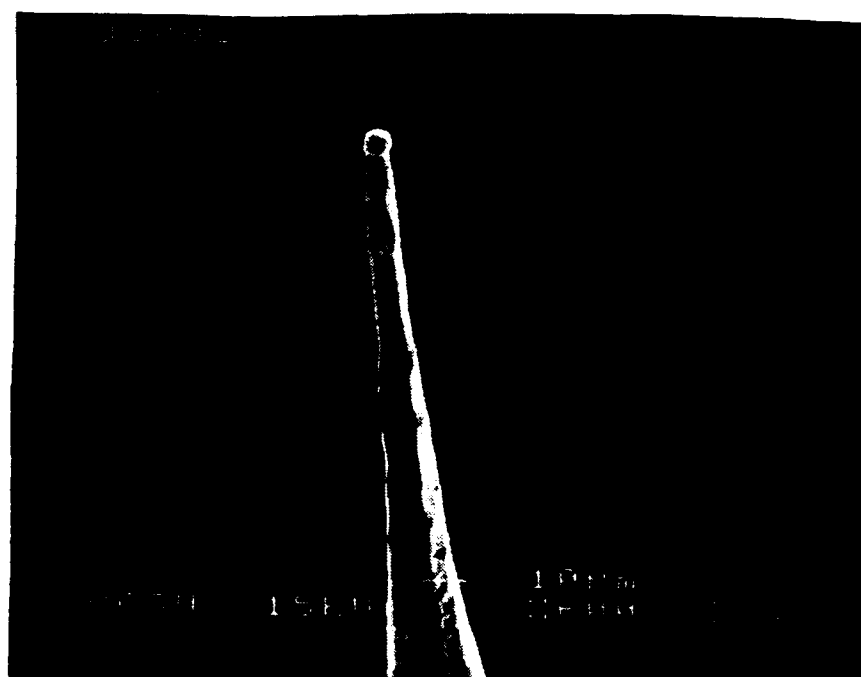
4.3.2 Automated Polishing Process Sequence

After bulk etching, a rinse step is necessary to remove solids deposited from the solution by the vigorous bulk etching process and any liquids adhering to the bulk etched wire. The bulk etched wire is inserted 3 mm into the solution, etched for 4 seconds at 1.5VAC, and then withdrawn.

To begin the polishing sequence (illustrated in Figure 3), the 1.5VAC polishing potential is turned on and the bulk etched tip is moved toward the solution. The position of first contact, denoted as position P1 (Figure 3(a)) is then detected by the computer and stored. The movement of the tip into the solution is continued and the position at which the current increases to 10mA (P2, Figure 3(b)) is also detected and stored and the 1.5VAC potential is switched off. This sequence allows



(a) Bulk Etched tip at 150X magnification



(b) Bulk Etched tip at 600X magnification

Figure 2: Bulk Etched Pt/Ir tip

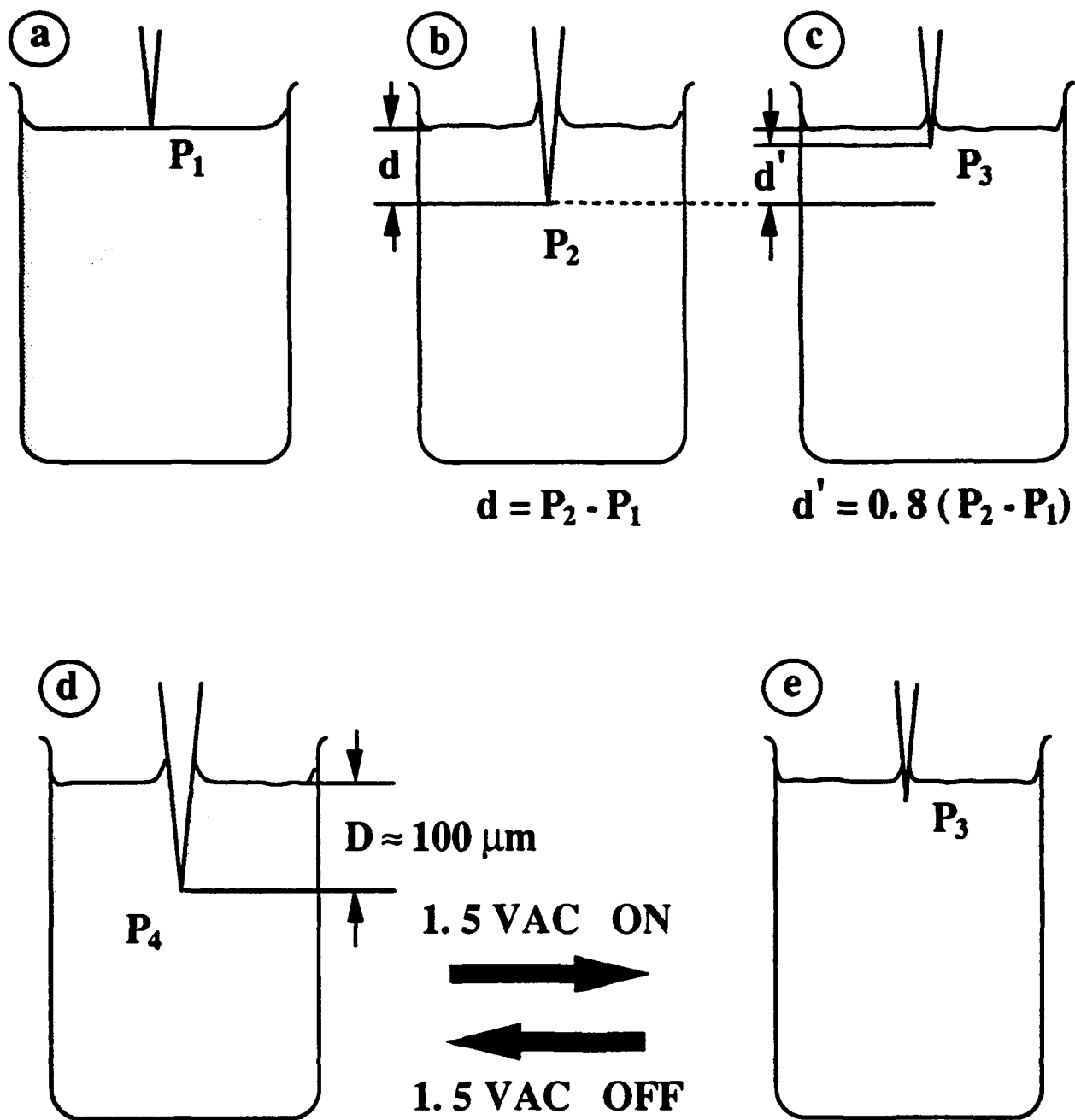


Figure 3: Schematic diagram of the position sensing for automated polishing sequence (a,b,c) and the polishing sequence (d,e).

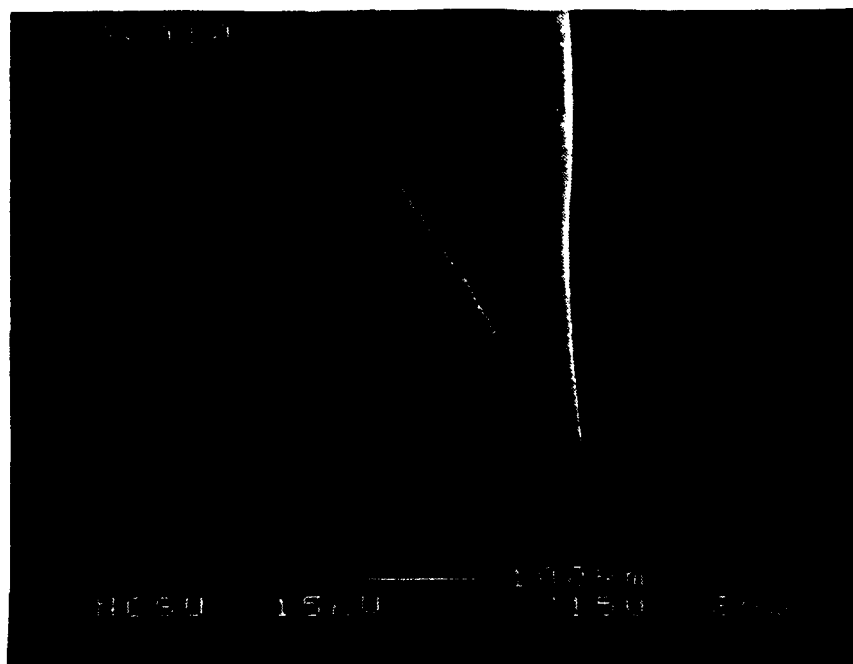
automatic detection of the apex of the bulk etched tip and the point at which the capillary action of the solution draws the solution up to the point where the bulk etched tip begins to thicken. The tip is then withdrawn by 80% of the distance between P1 and P2 (Figure 3(c)) or 100 μ m, whichever is less, to position P3 thus positioning the tip such that any excessive length will be polished off.

Once the cleaning and position sensing sequence is complete, the actual polishing sequence can begin. The polishing sequence begins by inserting the tip 100 μ m into the solution to position P4 (Figure 3(d)). The 1.5VAC polishing potential is then switched on and the tip is retracted to P3 (Figure 3(e)). This insertion and retraction sequence is repeated (Figures 3(d), 3(e)) with the etching potential applied only during retraction. The polishing sequence is repeated until the etching current diminishes to 0.024mA indicating that the unwanted portion of the tip has been etched away. As the tip is retracted, the area of the tip available for adhesion of the solution's capillary forces is reduced in a non linear fashion causing the falling away of the solution from the tip to accelerate as the tip is retracted. This additional polishing action insures that the residence time in solution of a very fine tip is small thus reducing the likelihood of over polishing. When the current has reached the desired minimum level, a few more polishing steps are added to remove any microscopic remains of the portion of the tip that was in solution. The tip is then removed from the apparatus, rinsed in a nitric acid solution (1ml HNO₃/ 50ml H₂O), and stored for later use. A typical example of a polished tip is given in Figure 4.

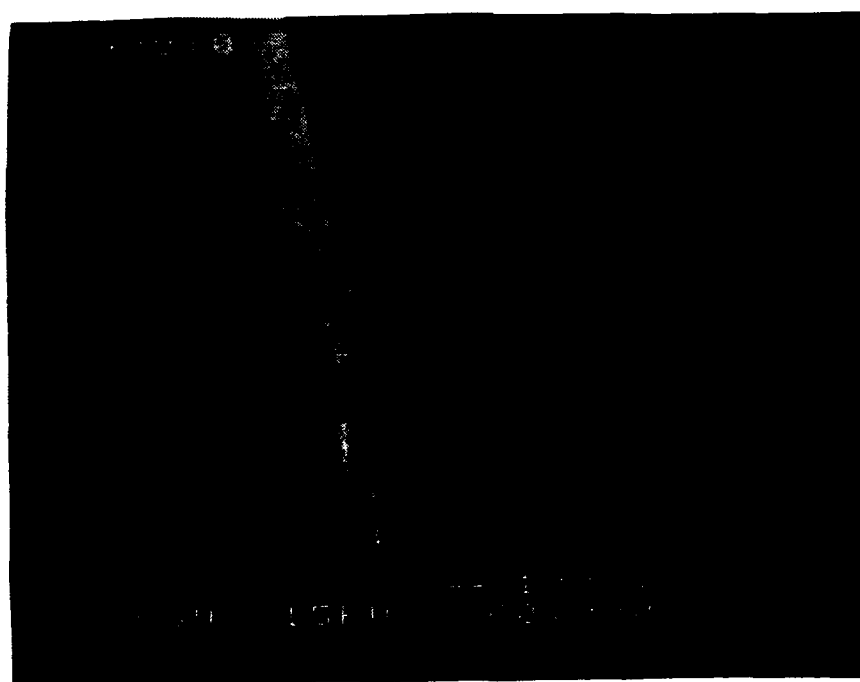
4.2 RESULTS AND DISCUSSION

Tips produced by the procedure described above were examined by electron microscopy to determine their quality. The criteria for a good tip at this stage of development is that the diameter of the apex must be less than 50nm. Based on this criteria, the process to date is able to achieve a 50% success rate under full automation. Average time per tip is 4 minutes per bulk etch and 10 minutes per polish giving a total for the entire automated process of 14 minutes. Adding one minute for rinsing and storage gives a production of 160 tips per week if the process were 100% efficient. At 90% efficiency, two such systems could, with minimum human intervention and judgement, produce approximately 250 tips per week, a number similar to the current production at MAS. A single individual should easily operate up to 5 automated system simultaneously raising production to more than 600 tips per week.

To raise the efficiency of the process to an acceptable limit i.e. greater than 90% good tips, a more detailed analysis of the interaction of the tip with the surface of the solution is necessary. The bulk etching process appears to be sufficiently well understood to produce the tip geometries necessary for automated polishing. It has been determined that the proper bulk geometry is produced using 27.5VAC with a tip immersion of 1.5 mm and that the endpoint can be automatically detected to very high efficiency.



(a) Polished tip at 150X magnification



(b) Polished tip at 60,000 magnification

Figure 4: Polished Pt/Ir tip

The cause of the low success rate is due to the polishing step and appears to be a result of the inability to predict the degree to which the solution will be pulled up onto the bulk etched tip with the 1.5VAC polishing potential applied. It has been determined that the operating point with respect to the edges of the solution container is a very important factor with respect to the above. Operating near the center of the solution during the polishing step produces bad tips due to difficulty in detection of P1 and P2 (described in Figure 2), caused by the aggregation of bubbles produced around the tip during the polish etch. The tip must be positioned sufficiently near the edge of the container so that the slope of the solution caused by capillary attraction to the edge of the container is such that the bubbles do not aggregate around the tip but rather flow upward toward the edge of the container. It remains to be determined if it is possible to position the tip to achieve reproducible results. It may be necessary to produce a container which forces positioning of the tip in a particular place or to modify the properties of the solution by addition of a surfactant to the solution.

4.3 CONCLUSION

The bulk etching process is sufficiently understood to reproducibly produce the desired results. Problems however appear in the polishing process. Tips having 50nm or less tip radii can be produced using the fully automated tip production procedure (bulk etch + polishing) but with a success rate of approximately 50%. Further work is required to raise this success rate to an acceptable level.

Reference

1. Musselman, I.H. and Russell, P.E., "Method of Fabricating Scanning Tunneling Microscope Tips", *U.S. Patent # 5,085,746*.
2. Musselman, I.H. and P.E. Russell, "Pt/Ir Tips With Controlled Geometry For STM", *Precision Engineering Annual Report*, pp. 299-314, 1990.

5 LINEAR SLIDE ACTUATORS FOR LONG RANGE MOTION WITH NANOMETER ACCURACY

James F. Cuttino

Graduate Student

Thomas A. Dow

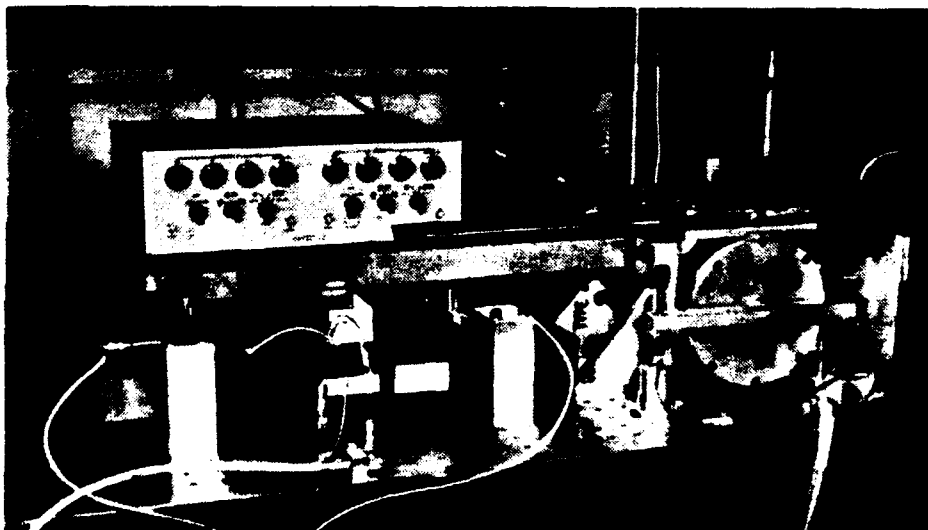
Professor

Department of Mechanical and Aerospace Engineering

A Precision Linear Optimization Testbed (PLOT) has been developed to study the ability of different actuator systems to provide long range motion with nanometer accuracy. Three actuators have been chosen for the study including a ball screw, a traction drive, and a piezoelectric inch worm. Characterization of the actuator performances is the first step toward controlling these systems to produce the desired accuracy in motion.

Nonlinear characteristics of the ball screw have been identified using a small angle rotation fixture (ARC) to input small amplitude angular displacements. Relationships between torque input to the ball screw and the linearity of the slideway displacement were experimentally measured and an analytical model was developed to explain the nonlinearities.

Preliminary tests have also been conducted using a linear traction drive. These experiments indicate that the traction drive provides a much more linear response than the ball screw, but some drift is apparent which may be due to the aerostatic bearings in the drive and any fluctuations in air supply.



5.1 INTRODUCTION

High precision machinery such as diamond turning machines and step and repeat cameras are responsible for the production of increasingly accurate products including optical assemblies, semiconductors, and micromachines. In many cases linear slide mechanisms play a crucial role in the accuracy of such machinery [1]. Much interest has therefore been directed toward the development and characterization of linear slides and their actuators. A Precision Linear Optimization Testbed (PLOT) [2] has been developed to examine the performance characteristics of three types of actuators; a ball screw, a traction drive, and a piezoelectric inch worm. The identification of response characteristics for these drives are important for accomplishing controlled motion with nanometer resolution.

Experimental results have shown the response of a ball screw to be nonlinear for motions less than 1200 nm, meaning that classical approaches to design problems do not necessarily apply to motion of such small magnitudes [3]. Several studies have been conducted dealing with conventional uses for ball screws with emphasis on high speed, wear, and preloading [4-8]. Others deal with general contact of two elastic bodies and the normal and tangential forces between them [9-12]. However, none of these specifically addressed the issue of nonlinearities at small displacements.

One particularly useful study by Ishikawa and Suda [13] discusses the introduction of friction to a ball in a V-shaped groove. These authors suggest that the friction is caused by differential slipping taking place in the contact area between the ball and the groove. By modifying this theory and applying it to a ball rolling in the grooves of a ball screw, significant insight is gained as to the sources of nonlinearities when the ball undergoes nanometer motion. With this approach, sources of nonlinearities in the ball screw were identified and the analytical model presented here was developed.

Similar work has been started to identify the performance of a traction drive system connected to the PLOT. A major difference between the traction drive and the ball screw is the significantly different "pitch" of the two systems. The ball screw has a 5 mm lead meaning that it moves 5 mm per turn of the input shaft. The traction drive, however, has an apparent lead of 199.5 mm which is 40 times larger than the ball screw. This means that much smaller angular motions will be required for the same linear displacements. A fixture has been fabricated to induce small angular displacements on the order of 0.01 arcseconds to the traction drive. Preliminary results indicate that the response of the slide to angular input to the traction drive is much more linear than that for the ball screw. Unfortunately, little literature directly addressing traction drives is currently available. J. Kannel has developed an analytical model for the traction between cylinders with specific application to traction drives [14]. Much work has also been conducted on general applications of cylinders rolling on a flat surface. Extensive research has been conducted on

deformable bodies in rolling contact, especially in the study of rail road wheels which may be applied to traction drives [9, 15-18]. Studies on rolling friction are also available [10, 19].

5.2 BALL SCREW ACTUATOR

For the ball screw experiments, the testbed consisted of a ball screw, a linear aerostatic slideway, and a non-influencing coupling connecting the drive system to the slideway as shown in Figure 1. For these small displacement experiments, the screw was activated using an Arc-Second Rotation Fixture (ARC), which consisted of a flexure driven by either a piezoelectric actuator or a motorized micrometer. The PZT actuator was used for displacements of less than 100 nm and the motorized micrometer was used for those greater than 100 nm. The input angle was measured by a capacitance gage on the ARC fixture opposite the actuator. Slide displacement was measured using a pair of differential capacitance gages developed at Lawrence Livermore National Laboratories, which were calibrated with an ADE MicroSense 3401 capacitance gage shown in Figure 1. Torque applied to the fixture was measured using a load cell in line with the PZT and the motorized micrometer. The torque contribution from the ARC fixture was determined by loading the ARC with the ball screw removed and measuring the relationship between the applied torque and the fixture angle. This slope was then subtracted from the torque data for subsequent experiments.

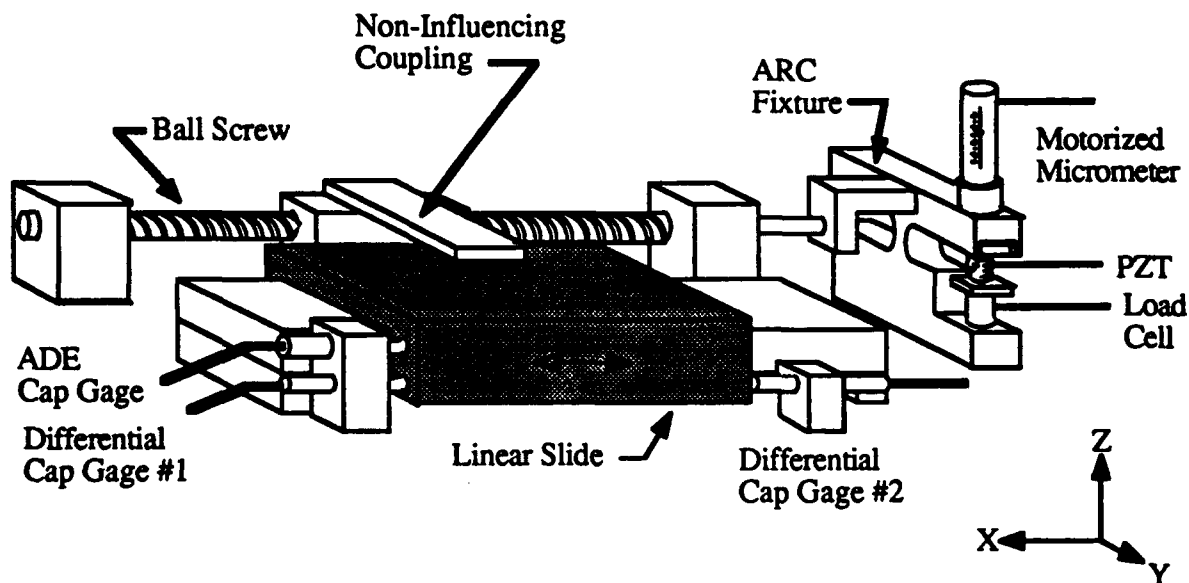


Figure 1: PLOT test fixture.

5.2.1 Displacement Response to Differing Input Excitations

Experiments were conducted to examine the ability of the PLOT to follow different types of inputs. These input signals included step input, ramp input, and sinusoidal input excitations. The step input experiment was conducted to examine the ability of the system to follow displacement commands of several nm. Figure 2 shows the slideway response to step inputs of approximately 3

nm over 2.5 second periods. The response of the system is quite impressive considering the open-loop nature of the system, with a noise level of approximately 2 nm for a mean step of 2.8 nm and a repeatability of less than 1 nm. Note that the input voltage is included for qualitative purposes only and has no relationship to the y-axis scale.

Increasing the displacement range of the experiments to over 100 nm gave the first indication that nonlinearities existed in the system. This is illustrated in the ramp input case shown in Figure 3. For an input generating a 120 nm displacement, a nonlinear response is readily apparent as the signal deviates from the dotted line in Figure 3. This deviation indicates hysteretic effects have been introduced into the system. The deviation from a linear response is as high as 20 nm, corresponding to 17% error, and the system displays a higher slope towards the end of the cycle and returns to zero net displacement.

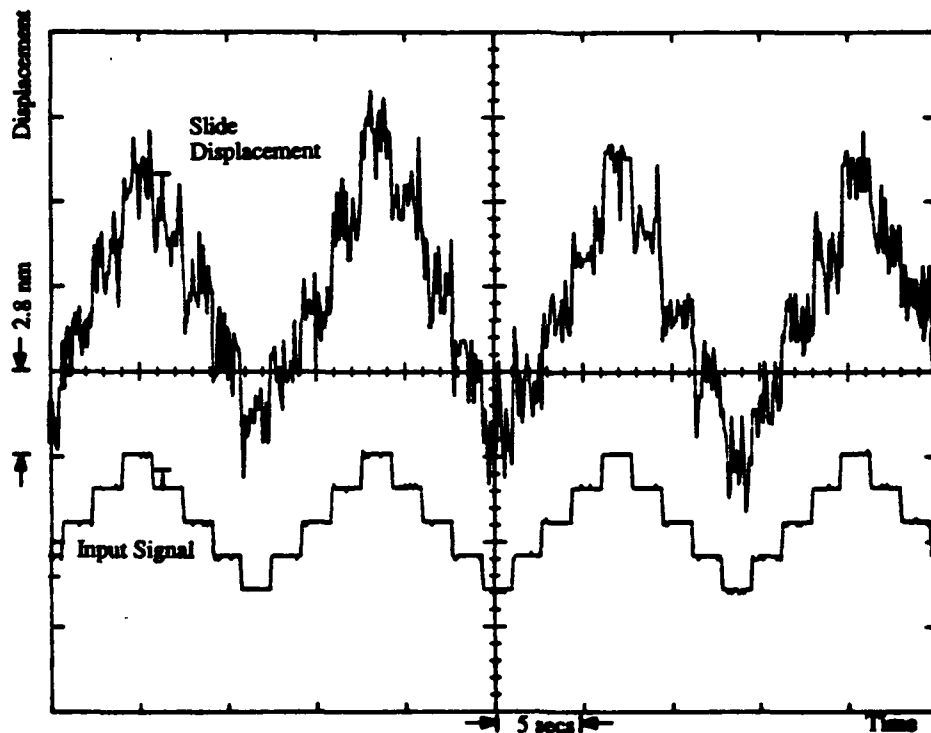


Figure 2. Slideway response to 3 nm input steps at 2.5 s intervals.

5.2.2 Hysteresis Curves

Figure 4 shows the measured displacement of the slideway as a function of the torque input. As increasing torque is applied to the ball-screw, the slideway displacement increases. The torque reaches a steady state value around 160 N-mm, signifying that all of the balls in the nut are rolling. The torque is then reduced and the ball-screw is driven in the opposite direction until the maximum torque is again reached. Note that the slope of the torque/displacement curve changes dramatically when the direction of motion is reversed.

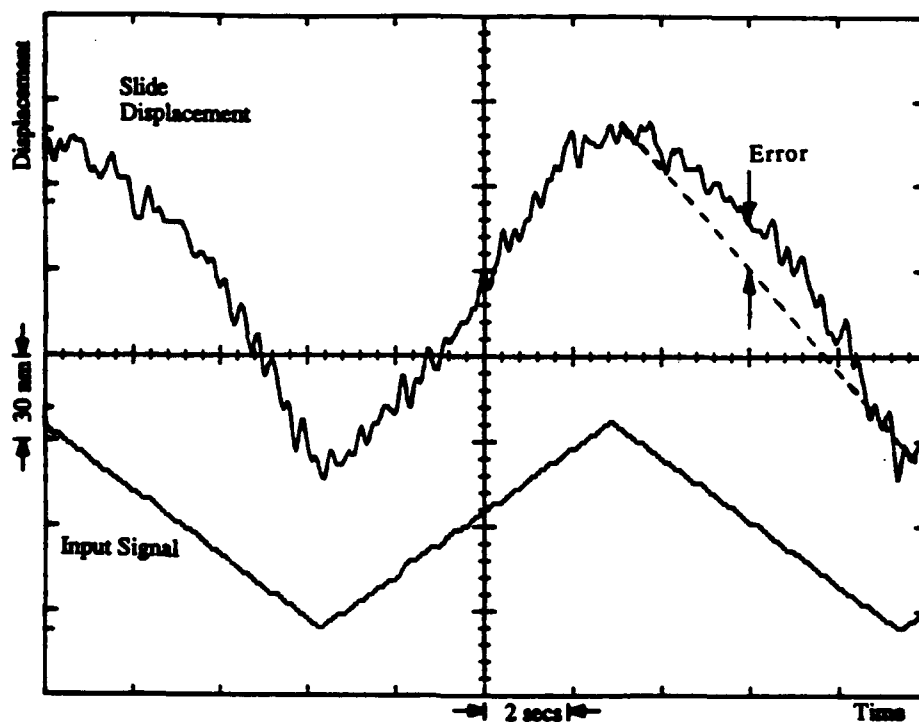


Figure 3: Displacement vs. Time Response for Ramp Input (120 nm ramp input over 12 s period).

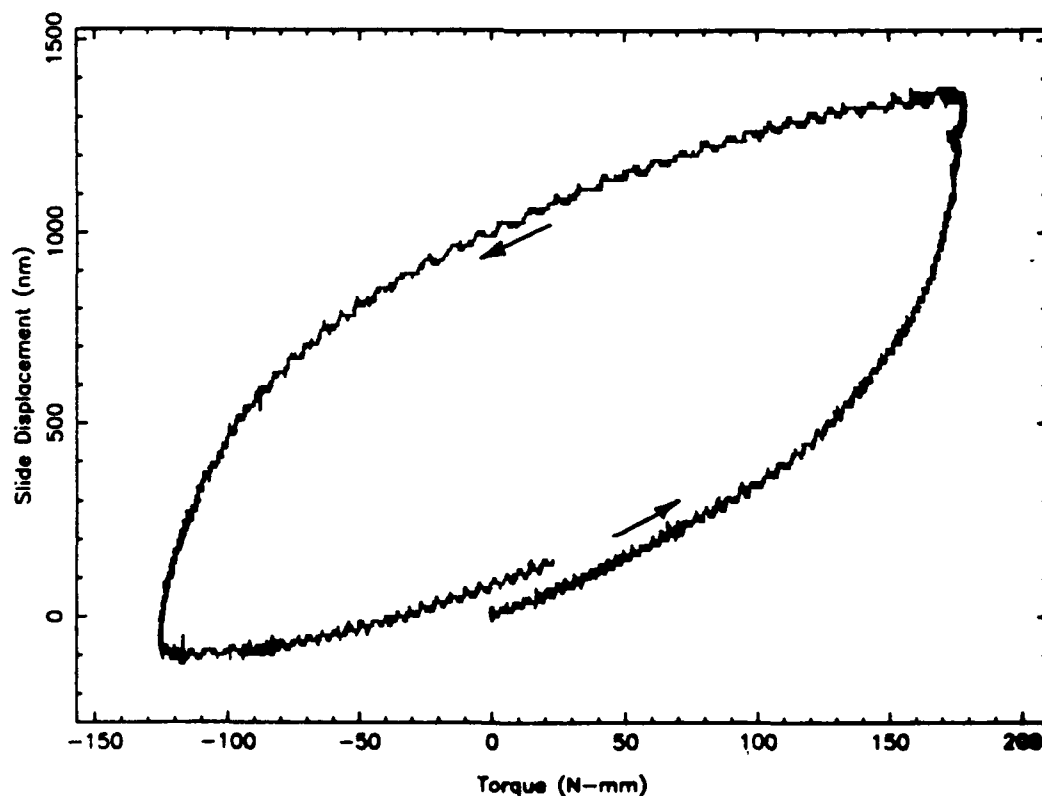


Figure 4. Ball-screw displacement vs. torque.

The shaft of the ball screw was machined as shown in Figure 5 to allow different mounting configurations. It was originally designed such that the ARC would drive the thrust bearing end (Side A) and then modified so that the ARC would drive the radial bearing end (Side B) with the radial bearing removed.

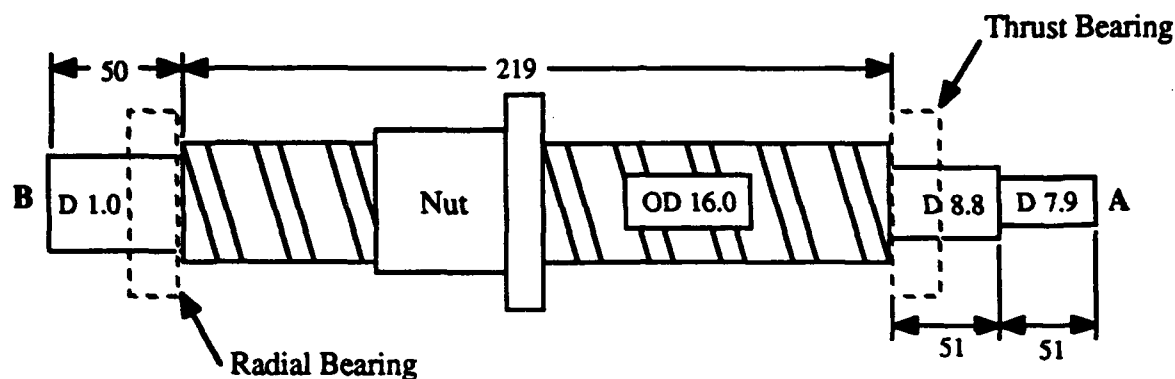


Figure 5. Screw Dimensions as Modified for PLOT (dimensions in mm).

Figure 6 shows the relationship between slide displacement and input angle when driving the system from the thrust bearing end. For a perfectly rigid system, the ratio of slide displacement to input angle should be equal to the specified pitch of the ball screw (3.92 nm/arcsec for the NSK W1603W-80P-C3Z5.08 ball screw used on the PLOT). However, Figure 6 illustrates a nonlinear relationship in the measured pitch of the ball-screw as a function of input angle. For small angles (less than 200 arcsecs), the effective pitch of the system is significantly less than that of the screw itself. For motion more than 200 arcsecs, when the torque reaches its steady-state value, the system motion follows the pitch of the screw.

The effect of screw cross section and therefore windup on the ball screw performance was studied by driving the screw from the radial bearing end (the results of which are shown in Figure 7) and comparing the results to those in Figure 6. Due to geometric constraints, the closest the slide could be positioned to the center of the screw was such that the nut was 147 mm from the ARC fixture when driven from Side A, and 109 mm when driven from Side B. The lines drawn on each side of the response curves were drawn to measure an "envelope" or width of the response and have a slope equivalent to the target pitch of the ball screw. Using these markers, the width of the response in Figure 6 is found to be 130 arcsecs, indicating a significant nonlinearity. The response shown in Figure 7, however, has a width of only 63 arcsecs, implying that the slide follows the input of the screw more closely when driven from Side B. Again, if the system were perfectly rigid, a linear response with a slope equal to the pitch of the screw would be expected.

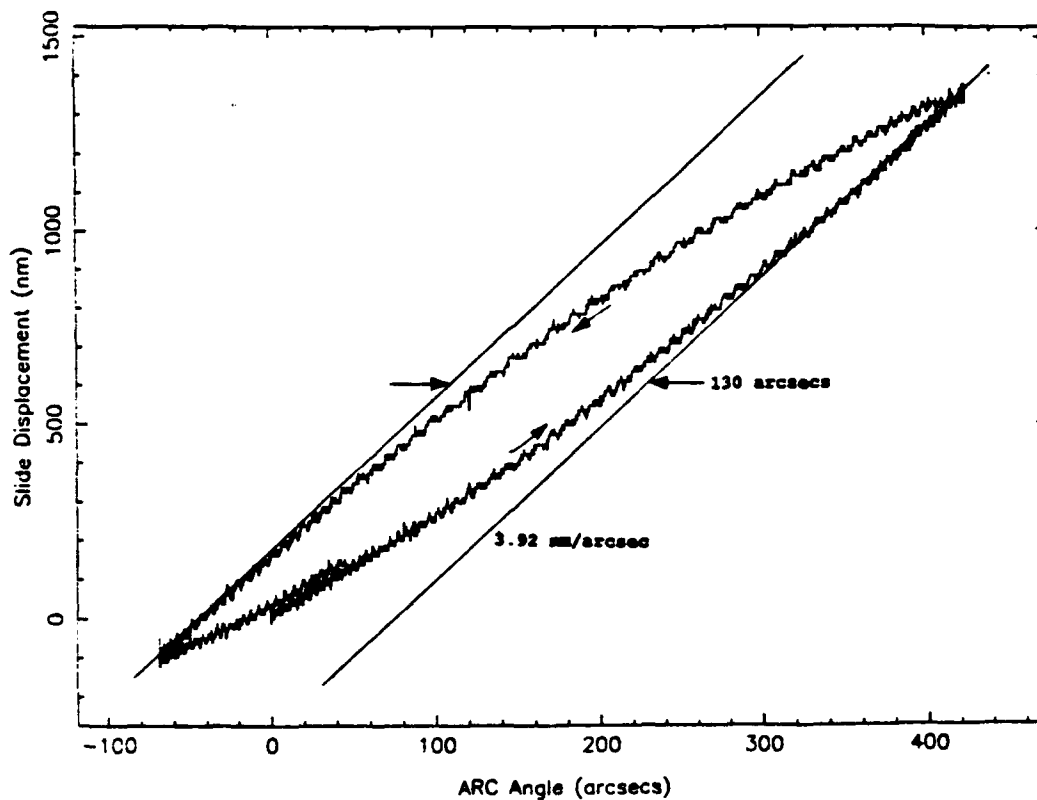


Figure 6. Ball-screw displacement vs. angular input when driven from Side A.

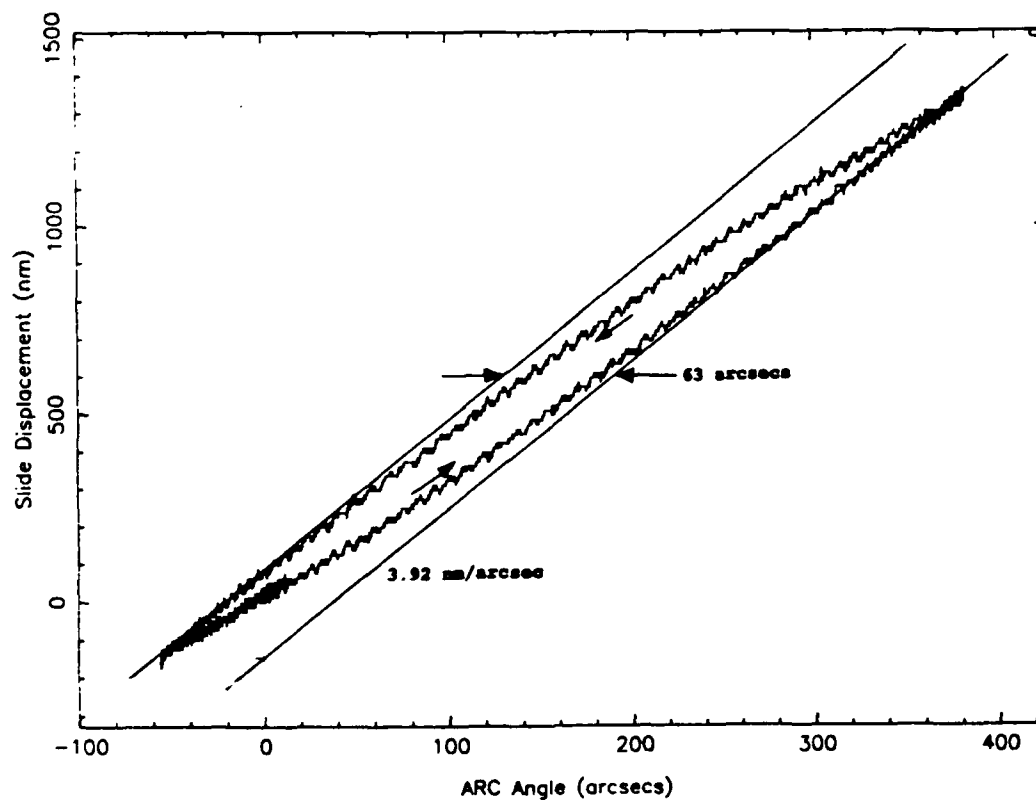


Figure 7. Ball-screw displacement vs. angular input when driven from Side B.

5.2.3 Model of ball-screw Friction

The change in system response with angular input is a problem for positioning slideways where sub-micrometer accuracy is required. In most instances, the ball-screw is assumed to be a simple gain in the control algorithm; that is, there is a fixed relationship between input angle and linear displacement. Figures 6 and 7 show that this is not true for displacements less than 200 nm and a model of this nonlinear behavior is the first step in developing a control system to improve the positioning accuracy of this type of actuator.

The steady-state torque required to rotate the ball screw is assumed to originate in the forces necessary to roll/slide the balls between the stationary nut and the revolving screw. A combination of the torsional deformation of the screw resulting from the ball interface forces can be shown to result in the nonlinear responses illustrated in Figures 4 and 6.

Ball Screw Geometry The interface pressure between the ball, nut, and screw is determined by the geometry of each component and the normal load. In the NSK ball screw used on the PLOT, the diameter of the ball is larger than the clearance between the screw and the nut. This arrangement serves to preload the balls and to prevent backlash due to motion reversal. To reduce the friction inherent in such a design, every other ball has a reduced diameter and serves as a rolling spacer. Therefore, of the 80 balls in the working channel of the nut, only 40 are loaded.

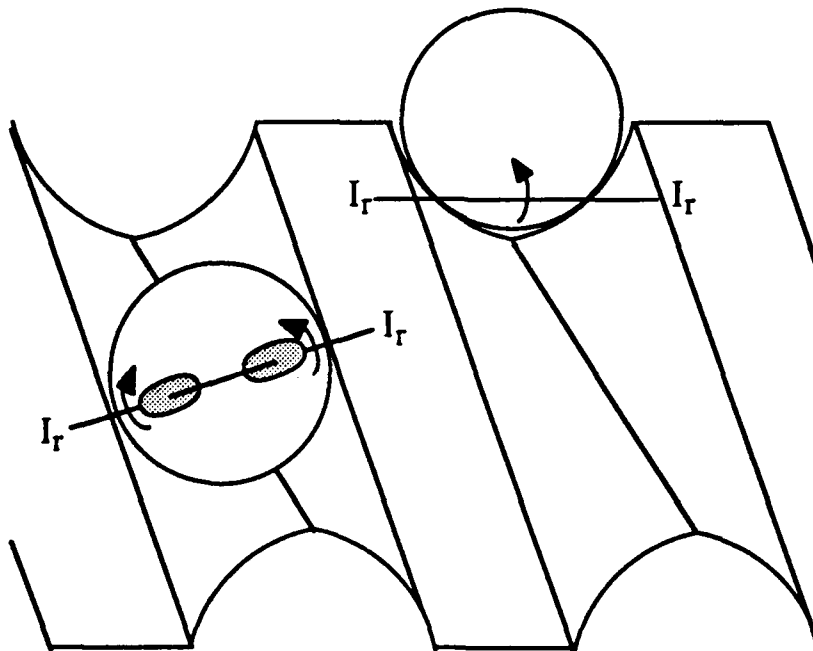


Figure 8. Contact between a ball and gothic arch groove in the ball-screw

The geometry of the ball screw contacts is shown in Figure 8. The screw and nut are formed with two large radii forming a "gothic arch". The large radius reduces the peak pressure at each of the elastic contacts by increasing the area of contact.

Studies conducted by Ishikawa and Suda [13] suggest that friction is introduced into a ball rolling in a V-shaped groove because portions of the contact between the ball and the groove are at different heights from the groove bottom and therefore velocities of these portions must be different, causing differential sliding.

The ball contact region over the gothic arch is calculated for this analysis using Hertzian theory and is shown in Figure 8. I_r is defined to be the instantaneous axes of rotation for the ball against the screw and is assumed to be the center of the contact patch. Since the ball rotates about an instantaneous center point, it is assumed that *any other point* in the contact patch must undergo rotation with respect to that center. In other words, only *one* point, the center of contact, does not undergo sliding. The angular rotation of the contact is the same as that of the ball.

Once the pressure distribution and contact patch shape are known, the total torque of the ball may be calculated based on a summation of the contribution of sliding friction from each point within the contact area.

Contact Patch Geometry The component geometry and preload were obtained from the manufacturer or, where proprietary (such as arch radius), were measured on the ball screw. The elastic contact areas between the ball and the nut and the ball and the screw were analyzed to find the pressure distributions and contact dimensions. These two contacts are different due to two factors. First, the contact with the nut is at a radius larger than that of the screw. Second, the curvature of the nut is negative, whereas the curvature of the screw is positive. This means that the contact between the ball and the nut is larger than the contact between the ball and the screw.

Torque Calculation Based on the pressure distribution and contact dimensions, the contact was divided into small areas, and, assuming a friction coefficient, the torque necessary to twist the contact about the instantaneous center of rotation was calculated and is shown in Figure 9. The torque is then given as

$$T = \mu r (P(x,y) dA) ,$$

where T is the torque, P is the pressure at (x,y) , dA is the element area, μ is the friction coefficient, and r is the distance of the element from the instantaneous center of rotation.

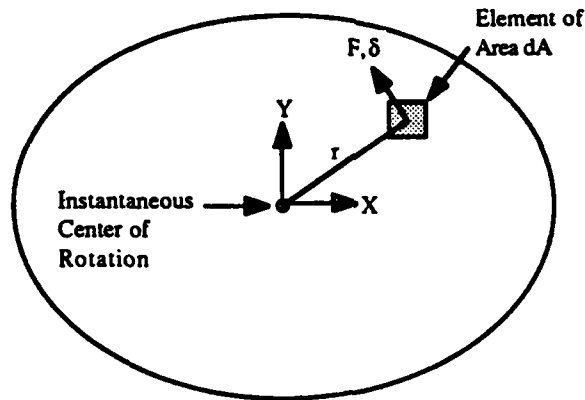


Figure 9. Contact Patch Discretization and Torque Calculation.

The pressure distribution is elliptical over the contact and is maximum at the instantaneous center of rotation as shown in Figure 10(a). The contribution of torque is zero at the center, increases away from center until a maximum torque is reached, and then decreases to zero again around the outer edge of the contact as illustrated in Figure 10(b). The calculated torque for a specified friction coefficient of 0.11 (representing lubricated steel sliding on steel at low speeds) is 1.06 N-mm for a

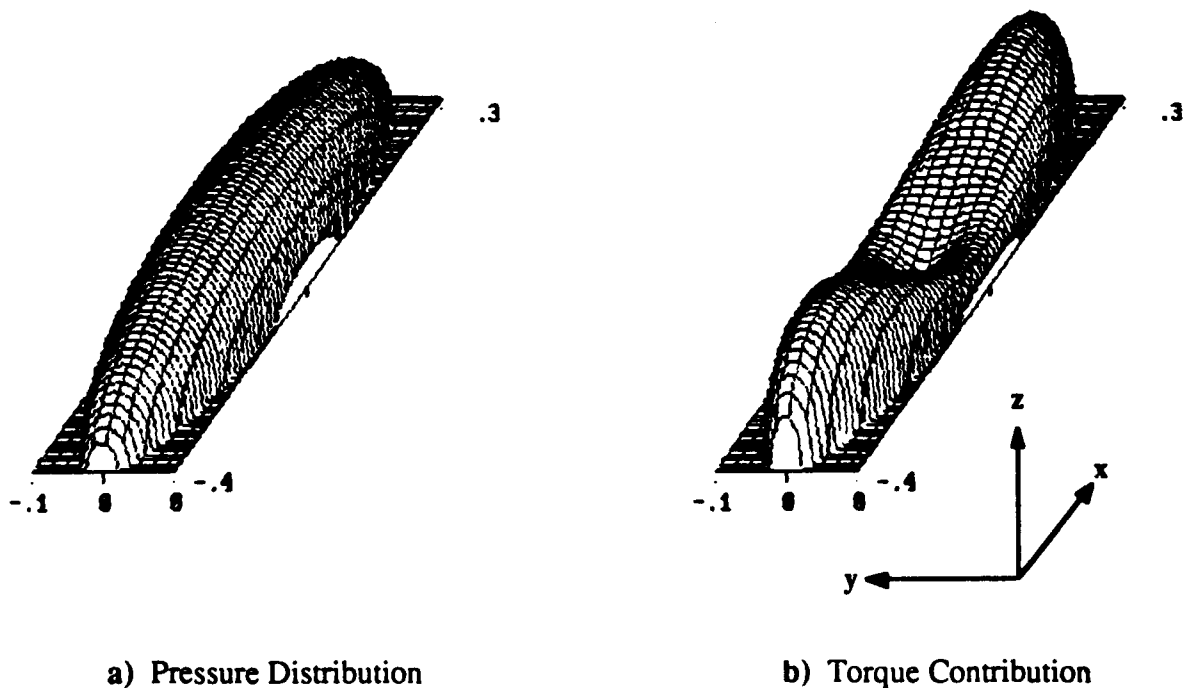


Figure 10. Analytical Pressure and Torque Distributions in Contact Patch.

single Ball/Nut interface and 1.07 N-mm for a Ball/Screw interface. Multiplying each value by 2 (2 contact patches per side) and adding yields a total torque of 4.25 N-mm per ball, or 170 N-mm for the 40 load-carrying balls in the ball screw. This value is very near the steady state torque measured for the PLOT apparatus.

Non-Linear Motion The friction model described above can also be used to determine the origin of the non-linear relationship between input angle and output displacement. The torque required to roll the balls along the groove between the nut and the screw will produce elastic deflections at the interface surfaces as well as in the screw itself. The torsional windup of the screw can be calculated from the screw cross-section, length, and magnitude of the torque. However, the interface between the ball and nut/screw is more difficult to model.

While it is impossible to exactly model the surface roughness of the ball or grooves, defining an array of *representative* elements and identifying an average stiffness can give significant insight into the response of the interface due to a load. In Figure 11 these elements are represented as columns for simplicity, but the roughness could be represented by cubes, cones, or pyramids. Each element is defined to have a stiffness, k , and a friction coefficient, μ . As the contact patch is rotated through an angle ϕ about the instantaneous center, each element is deformed a distance δ_i according to its distance from the center of rotation. The reaction force on that element is defined as

$$F_i = k \delta_i ,$$

and the contribution of that element to the torque on the contact patch is

$$T_i = F_i d_i$$

where d_i is the distance from the center of the contact patch. Each element is allowed to deflect until this resistive force equals or exceeds the frictional force

$$F_{fi} = \mu N_i ,$$

after which the element will slide and its deformation is kept constant. N_i is the normal force on that element as described by the Hertzian pressure distribution. At this point, the element's torque contribution becomes a constant value equal to

$$T_i = \mu N_i d_i .$$

The torque is multiplied by 4 to represent the four contact patches of the ball.

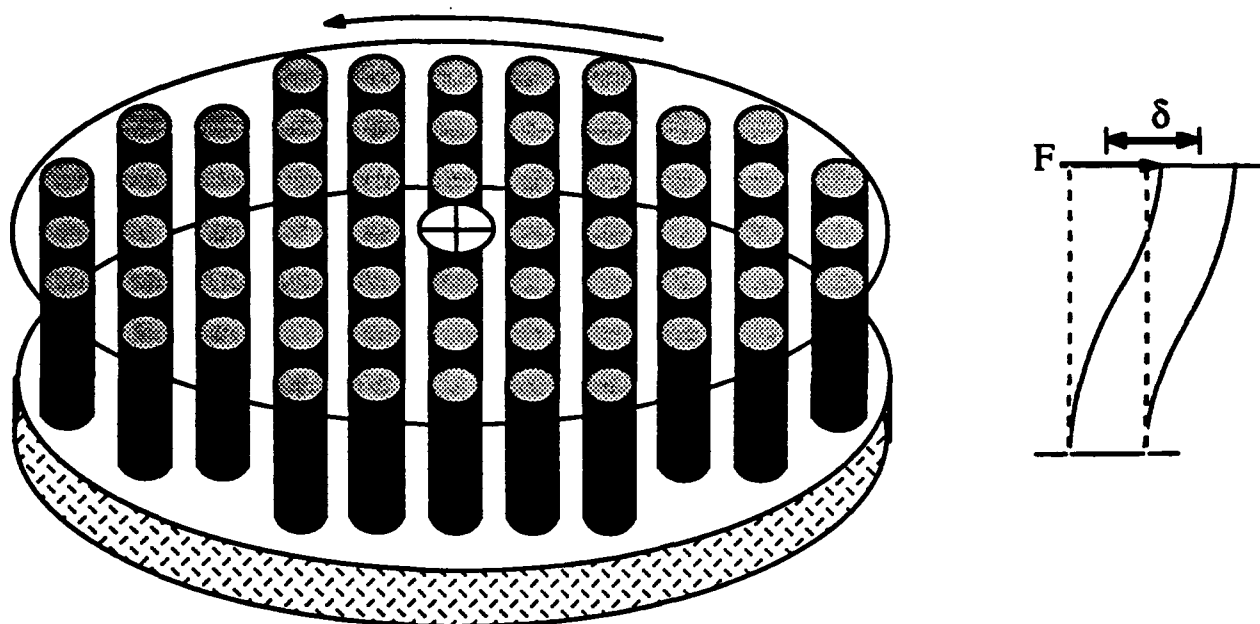


Figure 11. Modeling Surface Roughness in Contact Patch.

With this model, the outer elements will slide sooner than those near the center of rotation as a consequence of the longer moment arm and the small normal load at the outside edge of the contact. This will result in gradual breakaway and a nonlinear relationship between torque and rotation of the screw. This relationship coupled with the windup of the shaft can be used to predict the ball screw performance.

Once the relationships of torque vs. angular displacement of the ball were developed, they were coupled to the angular displacement of the screw using epicyclic analysis. This provided a relationship between ball rotation, ϕ , and screw rotation, θ .

Theoretical Results Two parameters are needed to model the performance of a ball-screw drive: the friction coefficient and the stiffness of the ball interface. The measured response was used to select these two values and resulted in a reasonable friction coefficient of 0.11 and an interface stiffness of 168 N/ μm .

The predicted performance of the ball-screw drive is shown as the dotted lines in Figures 12 and 13. The predicted curves start at the origin of torque and displacement and the experimental curves have been centered with respect to the predicted values. When the drive is first loaded, all of the interface elements are theoretically unloaded, a state which will never exist in practice. As the ball screw is loaded, however, some residual elastic deflection is stored in the interface and the response to subsequent motion is quite different as illustrated in these figures.

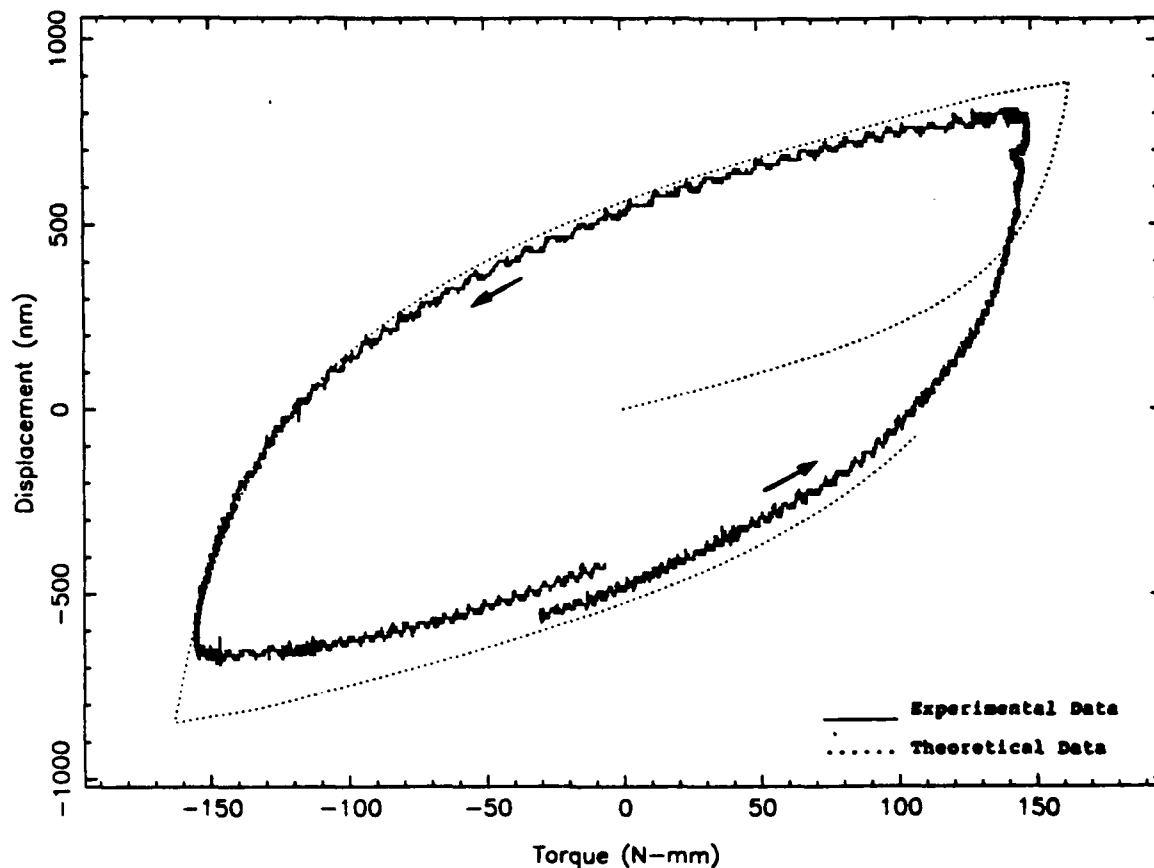
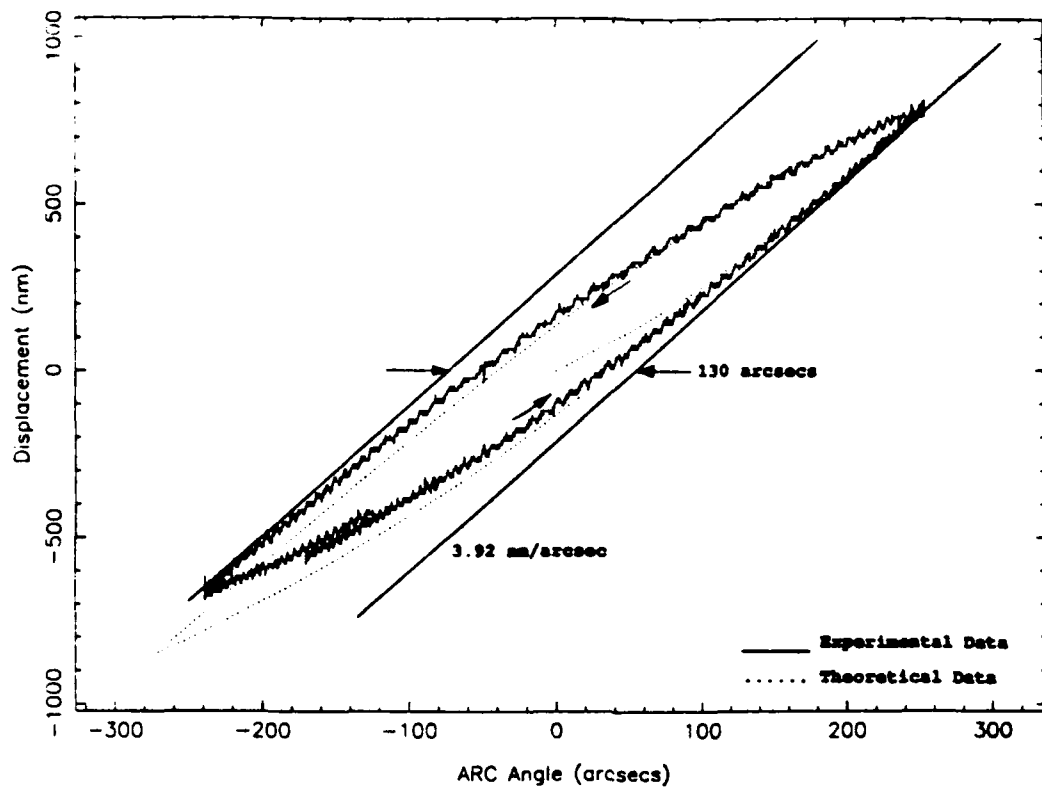


Figure 12. Ball screw displacement vs. torque.

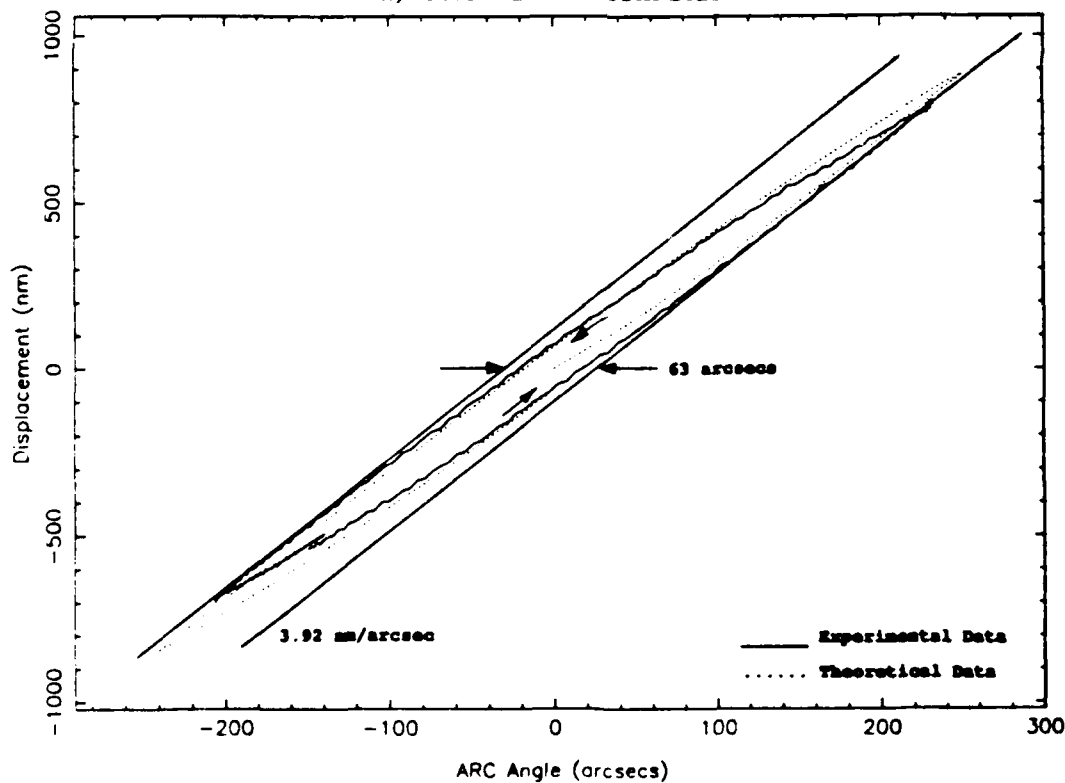
Figure 12 shows excellent agreement of the predicted values vs. the experimental results for displacement vs. torque. It can therefore be concluded that the nonlinear torque is due to the nonuniform breakaway at the ball contacts and subsequent sliding of the "high points" in the surface roughness.

The nonlinear torque response can also be used to explain the shape of the screw displacement curves of Figure 13. As the torque increases nonlinearly, a steady-state value of 150 N-mm is reached at a slide displacement of 700 nm. Since the windup is a function of the generated torque, it is also nonlinear until the displacement reaches 700 nm. As Figure 13(a) shows, it is at that point that the pitch equals that of the ball screw. The results of the model (shown as the dotted line) match the experimental curves well, including the width of the hysteresis curve (130 arcsecs) which is due to screw windup.

Figure 13(b) shows results for the model when driven from Side B. Here the width of the curve is 63 arcsecs which matches the model prediction. As shown in Figure 5, the difference in Figures 13(a) and 13(b) is due to the geometry of the screw. The small sections of the shaft near Side A provide significantly more windup than the larger sections near Side B. The model accounts for these differences and indicates that the hysteretic effects seen in the displacement vs. input angle plots is due to windup which is caused by the torque generated in the contact patches of the balls.



a) Screw driven from Side A



b) Screw driven from Side B

Figure 13. Ball screw displacement vs. angular input.

5.3 TRACTION DRIVE ACTUATOR

A traction drive manufactured by Rank Pneumo has recently been attached to the PLOT for performance tests. Few modifications to the PLOT were required to incorporate the new drive. The ball screw was removed and the traction drive was mounted to drive along the centerline of the slide as shown in Figure 14. New capacitance gage mounts were also fabricated to measure displacement along the same line as the bar/slide coupling. These steps were taken to minimize any yaw error in the slide motion. The traction drive has much travel clearance in the y-direction and is therefore not likely to effect straightness errors in that orientation. A flexure with z-direction mobility was used for coupling the slide, however, to minimize any errors due to vertical misalignment.

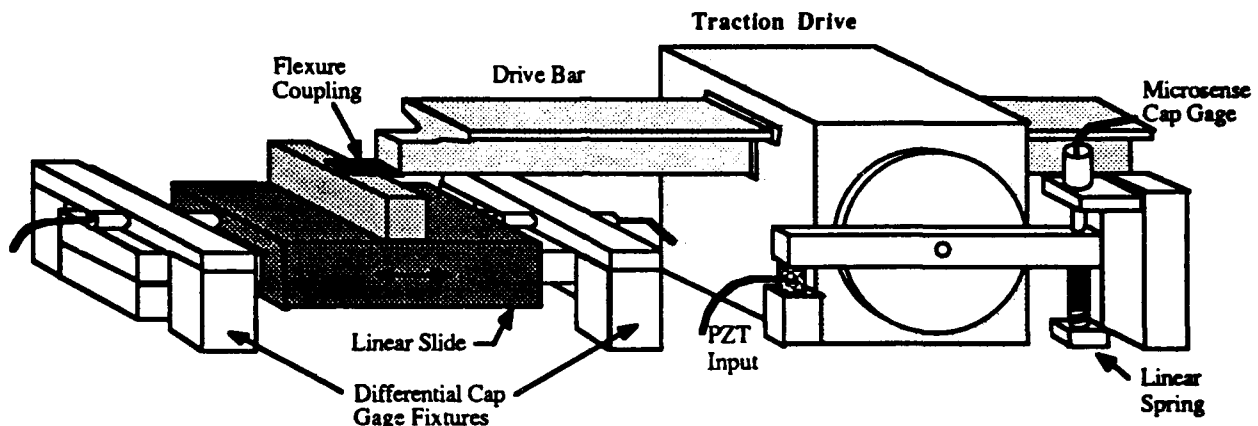
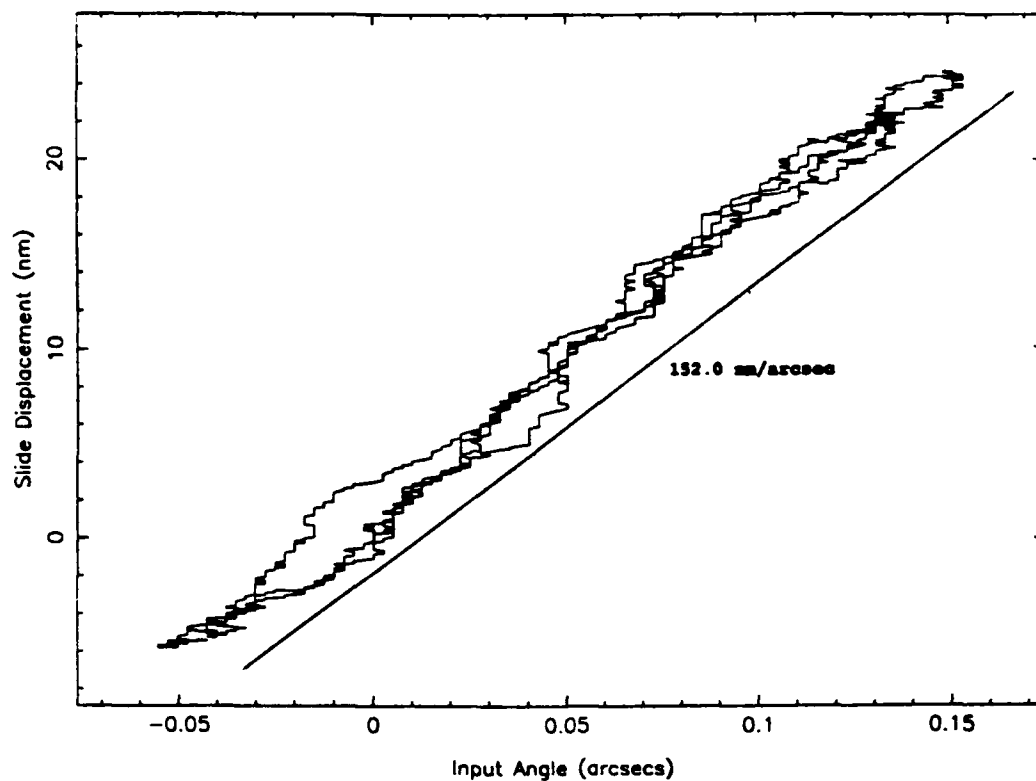


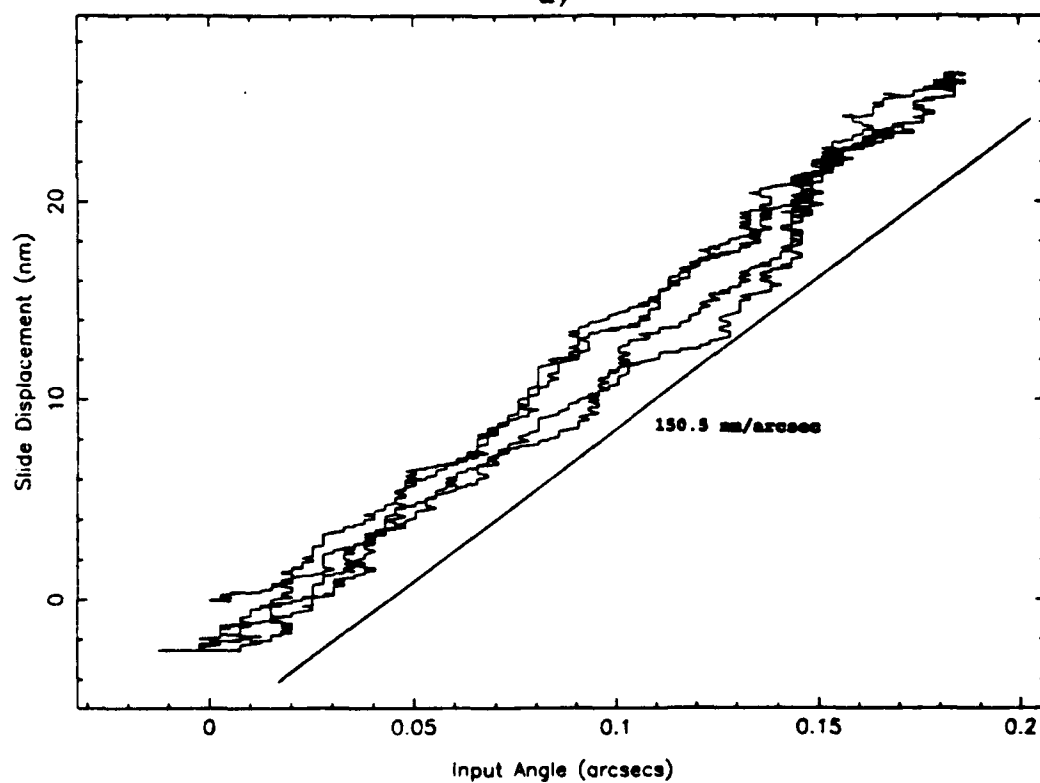
Figure 14. PLOT System with Traction Drive Actuator.

For preliminary tests, a cantilever arm was fabricated and attached to the drive to serve as a small angle rotation fixture similar to the ARC fixture used for the ball screw. This step was necessary for the traction drive since the ratio of linear displacement to angular input is 154 nm/arcsec , which is over 40 times larger than that of the ball ball screw. This large ratio, however, also meant that the range of the PZT was sufficient for displacements of several micrometers and the motorized micrometer was therefore unnecessary. For the preliminary tests, the PZT stack was excited with a sinusoidal signal (approximately 0.1 Hz) at differing input magnitudes to obtain the desired displacement ranges. Input angle was measured using the ADE MicroSense capacitance gage located over the cantilever arm at a nominal distance of 4.0 inches. The input displacement was opposed by a spring to reverse the motion of the drive.

Figures 15(a) and 15(b) show the results of the input angle/output displacement linearity tests for a 35 nm output range. Both plots were obtained under the same operating conditions at different times. No consistent overall curvature is evident in either graph, but the drift ranges as high as 4 nm over a single cycle. The plots seem to indicate that no appreciable hysteresis is present as in the



a)



b)

Figure 15. Slide Displacement vs. Input Angle for 30 nm range.

ball screw, but drift appears to be entering the system either through the slide, actuator, or capacitance gage readings. The average slopes are 151.96 and 150.51 nm/arcsec respectively, which are within 3% of the predicted value of 153.9 nm/arcsec.

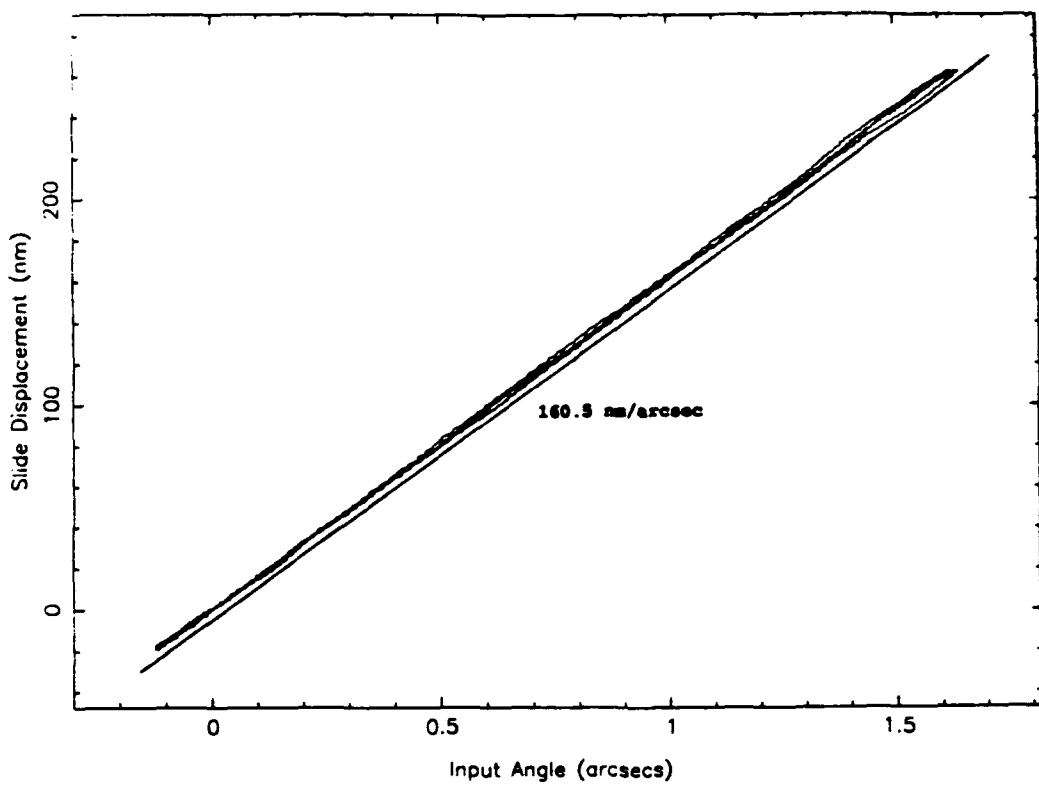
Figure 16 shows the results of two tests under similar conditions for a 300 nm range, or ten times the range of Figure 15. Again, the responses are linear with slopes of 160.5 and 161.1 nm/arcsec respectively and are within 2% of the predicted value of 153.9 nm/arcsec. Figure 16(a) shows very good repeatability with an overall drift of less than 8 nm, while Figure 16(b) shows a drift of as much as 15 nm over both cycles.

The sources for drift apparent in Figures 15 and 16 have not yet been identified. One possibility is that temperature fluctuations may be causing error in the experiments. Another possibility is that fluctuations in the air supply may be causing movement of the capstan itself. In the tests conducted on the PLOT, bottled air has been supplied to the slide to provide constant pressure. However, the traction drive is supplied by compressed air and there is a potential for pressure fluctuations when the compressor cycles. This hypothesis will be examined further in the upcoming weeks.

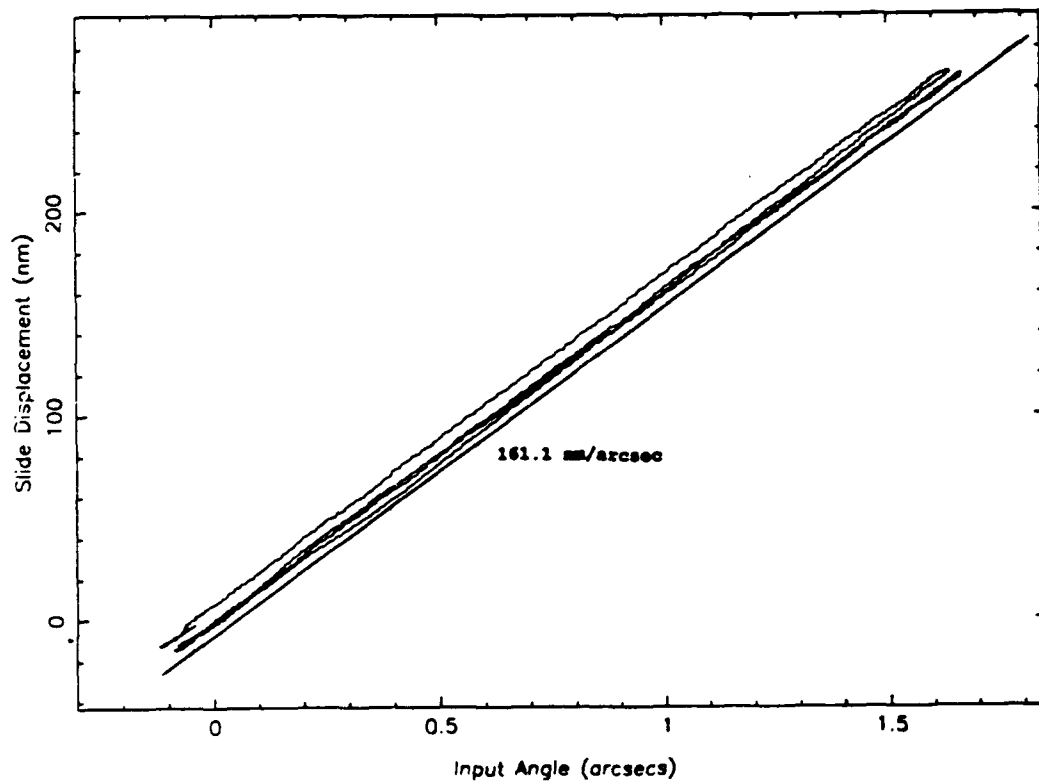
5.4 CONCLUSIONS

The preloaded ball-screw can be used for nanometer resolution actuation; however, the response is nonlinear. The nonlinearity in the torque has been modeled using discretized elements in the contact patch to simulate individual asperities on the surface. These elements are able to deflect based on a specified element stiffness until the generated resistive force exceeds the friction coefficient times the normal load, at which time sliding occurs and no additional torque is generated. The generation of this torque in turn results in windup in the shaft, which is the source in the nonlinearity in the slide displacement vs. input angle plots. A model has been derived which predicts the friction and displacement as a function of input angle. Characterization of this performance is necessary as the basis to develop a control strategy to operate over a wide range of displacement commands.

The traction drive appears to provide a more linear relationship between input angle and output displacement. However, sources of error do exist in the drift phenomena experienced in the tests. Disadvantages such as low stiffness and high output/input ratio can also make the drive difficult to control. Characterization of these parameters should provide valuable information for the design of control systems to produce nanometer resolution.



a)



b)

Figure 16. Slide Displacement vs. Input Angle for 300 nm range.

References

1. Futami, S., A. Furutani, and S. Yoshida, "Nanometer Positioning and Its Micro-dynamics", 1: p. 31-37, 1990.
2. Knight, B.F., *Development of a Testbed for Precision Linear Motion*. 1990, NC State University, Department of Mechanical and Aerospace Engineering:
3. Cuttino, J.F. and T.A. Dow, "Development of a Nanometer Motion Testbed" in Proceedings of Precision Engineering Center 1990 Annual Report, 1990. Department of Aerospace and Mechanical Engineering, North Carolina State University.
4. Belyaev, V.G. and R.K. Fattakhov, "Wear of Ball Screw Mechanisms", 8, n 1: p. 42-43, 1988.
5. McCarty, L.H., "Integral Spring Eliminates Backlash in Ball Screws", (February 15, 1988): p. 222-223, 1988.
6. Yeaple, F., "Rolling Replaces Grinding in Precision Ball Screw", (September 23, 1985): p. 108-109, 1985.
7. Belyayev, V.G. and R.K. Fattakhov, "Ball Bearing Screw Mechanisms of a New Design", 5, n 11 (November 1986): p. 5-7, 1986.
8. Weck, M. and T. Bispink, "Examination of High Precision Slow-motion Feed-drive Systems for the Sub-micrometre Range" in Proceedings of 6th International Precision Engineering Seminar, 1991. Braunschweig, Germany: Springer-Verlag.
9. Kumar, P., D.K. Sarkar, and S.C. Gupta, "Rolling Resistance of Elastic Wheels on Flat Surfaces", 126, n 2 (September 1, 1988): p. 117-129, 1988.
10. Domenech, A., T. Domenech, and J. Cebrian, "Introduction to the Study of Rolling Friction", 55 (3) (March 1987): p. 231-235, 1987.
11. Courtney-Pratt, J.S. and E. Eisner, "The Effect of a Tangential Force on the Contact of Metallic Bodies", 238: p. 529-550, 1957.
12. Hahn, G.T., K. Kim, and e. al., "Analysis of the Rolling Contact Residual Stresses and Cyclic Plastic Deformation of SAE 52100 Steel Ball Bearings", 109 (October, 1987): p. 618 - 626, 1987.

13. Ishikawa, Y. and M. Suda, "Differential Slipping Friction of a Ball Rolling in V-Shaped Groove", (n. 7): p. 29-34, 1986.
14. Kannel, J.W., *Development of an Analytical Model for Traction Between Cylinders*. 1985, North Carolina State University:
15. Kalker, J.J., "Transient Rolling Contact Phenomena", 14: p. 177-184, 1971.
16. Carter, F.W., "On the Action of a Locomotive Driving Wheel" in *Proceedings of Proceedings of the Royal Society of London*, 1926.
17. Kalker, J.J., "Rolling with Slip and Spin in the Presence of Dry Friction", 9: p. 20-38, 1966.
18. Kalker, J.J., "Three-Dimensional Elastic Bodies in Rolling Contact", Dordrecht, Boston, London: Kluwer Academic Publishers, 1990.
19. Bowden, F.P. and D. Tabor, "Friction - An Introduction to Tribology", Malabar, Florida: Robert E. Krieger Publishing Company, 1982.

6 DTM METROLOGY

Peter I. Hubbel

Graduate Student

Department of Electrical and Computer Engineering

G. McDonald Moorefield, II

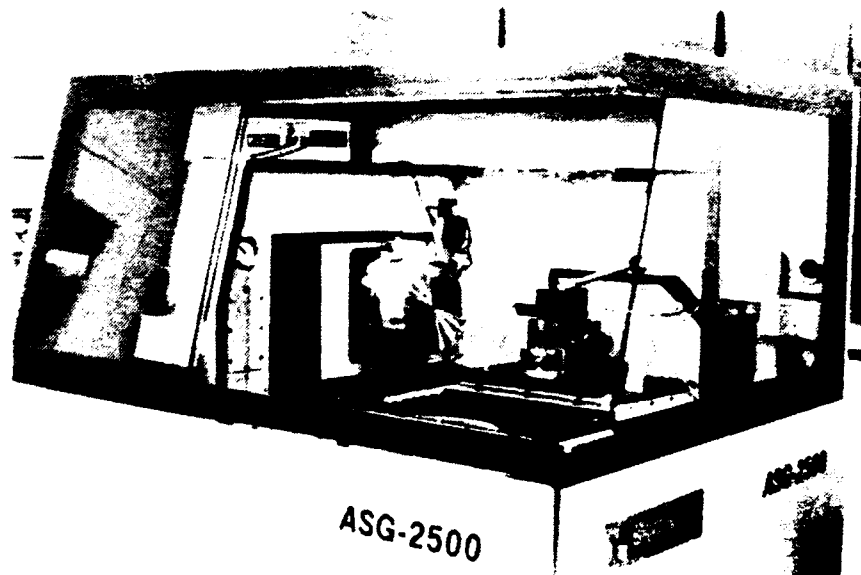
Research Assistant/Lecturer

Thomas A. Dow

Professor

Department of Mechanical and Aerospace Engineering

The identification and measurement of errors affecting the performance of the Center's Diamond Turning Machine (DTM) have continued. In previous reports, the error sources addressed have ranged from environmentally-induced laser interferometer error to slideway straightness error. In keeping with the error budget approach (which seeks to account for environmental, machine, workpiece, and operator errors), research efforts have included the measurement of axes positioning errors and spindle growth, an investigation to determine the cause of and identify a remedy for laboratory air temperature fluctuations, and an investigation into the role of the workpiece material in the tool centering process.



6.1 INTRODUCTION

The goal of the research in DTM Metrology is to assess the PEC's diamond turning (and grinding) capability as a precision fabrication process by identifying and evaluating sources of error. Because the fabrication process is influenced by the machine, the environment, and the workpiece, the sources of error range from straightness-of-motion error to environmentally-induced laser interferometer error to the intricacies of the material removal process. In the area of machine errors, recent efforts have focused on the characterization of positioning errors along the X- and Z-axes and the thermal growth of the workpiece spindle. In addition to these machine errors, an experiment is being drafted to explain the influence of the workpiece material on the tool centering process. Finally, in an effort to minimize environmentally-induced errors, a project was conducted to determine the cause of and implement a remedy for the laboratory air temperature fluctuations.

6.2 DTM AXES POSITIONING TESTS

Using the method outlined by Tlustý [1], positioning errors for the X- and Z-axes were measured. With this method, the positioning error was divided into dead zone error and scatter error (see Figure 1). Dead zone error is defined as the difference between the mean of the approaches from the positive direction and the mean of the approaches from the negative direction. Scatter error is \pm three times the standard deviation of the values from their respective mean values (i.e., with consideration of their direction of approach and their nominal point).

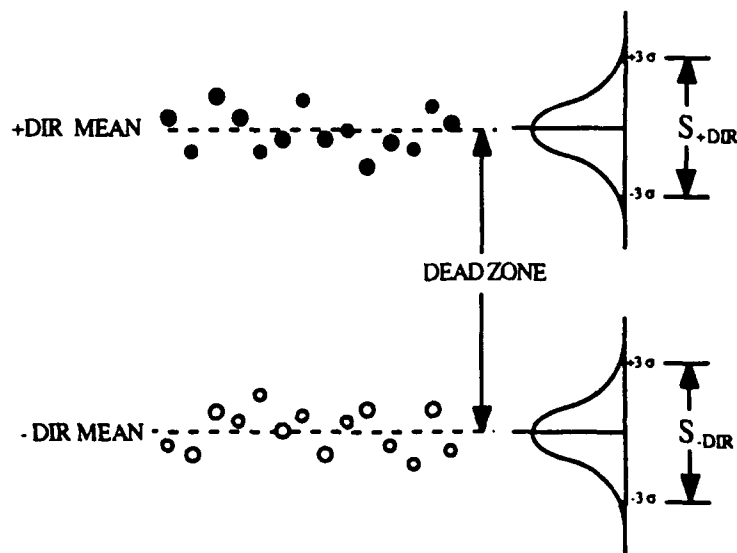


Figure 1: Dead Zone and Scatter Errors.

For each axis, the positioning error was measured when moving to two specified points (one point nominally at $L/4$ and another point nominally at $3L/4$, where "L" is the length of axis travel) from both the positive and negative directions (see Figure 2). For each point, forty approaches (twenty at 125 mm/min and twenty at 12.5 mm/min) were made from each direction. To allow positioning approaches from 5 millimeters away in both directions, the DTM's Zygo Axiom 2/20 linear laser interferometer system was used for the axis position measurements. (This technique of using the machine's laser interferometer system to measure the machine's positioning errors is not the ideal setup. However, this method was selected because of the ease of data collection. Furthermore, the errors in the machine's laser interferometer system were addressed earlier.) The DTM's controller automatically logged the position values into a datafile, sampling the laser interferometer data every millisecond. To filter out the vibrations of the slides (which have been separately identified), each one thousand (1000) laser interferometer values were averaged to yield one (1) datapoint for each one (1) second interval.

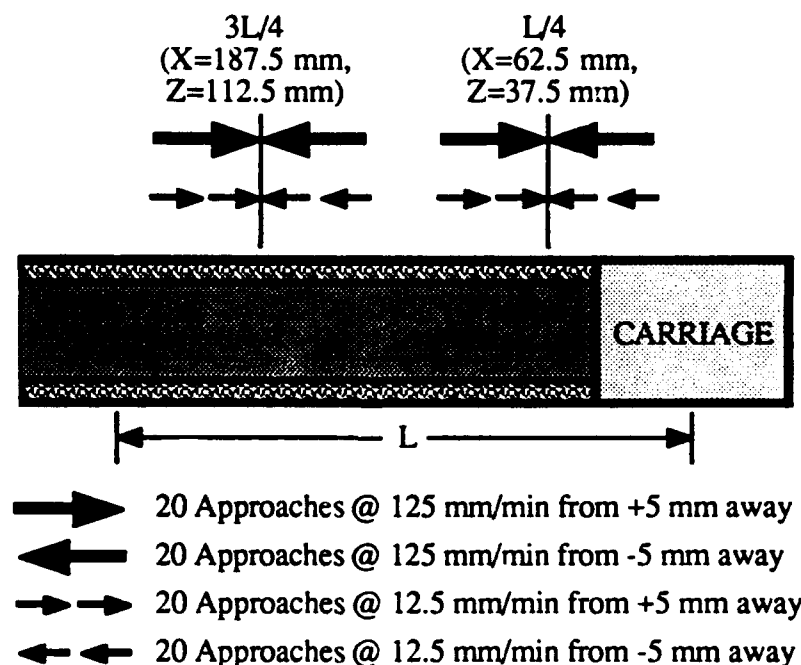


Figure 2: Procedure for Axes Positioning Tests.

The dead zone error (DZ) was calculated for the $L/4$ and $3L/4$ positions for each axis using the following equation.

$$DZ = \left| \sum_{i=1}^{n=20} \frac{\delta_{pos_i}(X)}{n} - \sum_{i=1}^{n=20} \frac{\delta_{neg_i}(X)}{n} \right| \quad (1)$$

The L/4 and 3L/4 dead zone values were then averaged to obtain an overall dead zone error for each axis.

$$DZ = \frac{DZ_1 + DZ_2}{2} \quad (2)$$

The scatter error (S) was calculated for each axis (with consideration of the direction of approach and the nominal point) using Equations 3-7.

$$A = \sum_{i=1}^{n=20} [\delta_{ip1,dir}(X) - \delta_{meanp1,dir}(X)]^2 \quad (3)$$

$$B = \sum_{i=1}^{n=20} [\delta_{ip1,-dir}(X) - \delta_{meanp1,-dir}(X)]^2 \quad (4)$$

$$C = \sum_{i=1}^{n=20} [\delta_{ip2,dir}(X) - \delta_{meanp2,dir}(X)]^2 \quad (5)$$

$$D = \sum_{i=1}^{n=20} [\delta_{ip2,-dir}(X) - \delta_{meanp2,-dir}(X)]^2 \quad (6)$$

$$S = \pm 3 \sqrt{\frac{A+B+C+D}{n_A + n_B + n_C + n_D}} \quad (7)$$

The total positioning error (TPE) was calculated as the sum of the dead zone and scatter errors.

$$TPE = DZ + 2S \quad (8)$$

The results of the positioning tests with the DTM's laser interferometer system are shown in Tables 1, 2, and 3. The individual positioning tests revealed that the Dead Zone, Scatter, and Total Positioning Errors for the Z-axis were smaller than those errors for the X-axis. However, when the results from the individual tests were grouped together (i.e., no longer segregating the data according to nominal position and speed), the overall results showed that Total Positioning Error for the Z-axis had approached the magnitude of the X-axis value. Noting that the data consolidation had essentially no effect on the Z-axis Scatter Error (the data consolidation had little effect on any of the X-axis error values), the cause of the increase in the Z-axis Total Positioning Error can be attributed to the increased Z-axis Dead Zone Error.

Position \Rightarrow	X=62.5 mm			X=187.5 mm		
Speed \Downarrow	Dead Zone (nm)	Scatter (nm)	Total Positioning Error (nm)	Dead Zone (nm)	Scatter (nm)	Total Positioning Error (nm)
Fast:	23	+/-13	49	19	+/-16	51
Slow:	18	+/-18	54	20	+/-18	56
Combined:	26	+/-16	58	25	+/-17	59

Table 1: Results of individual X-axis positioning tests.

Position \Rightarrow	Z=37.5 mm			Z=112.5 mm		
Speed \Downarrow	Dead Zone (nm)	Scatter (nm)	Total Positioning Error (nm)	Dead Zone (nm)	Scatter (nm)	Total Positioning Error (nm)
Fast:	15	+/-14	43	3	+/-12	27
Slow:	9	+/- 9	27	12	+/-9	30
Combined:	27	+/-12	51	34	+/-10	54

Table 2: Results of individual Z-axis positioning tests.

Axis \Downarrow	Dead Zone (nm)	Scatter (nm)	Total Positioning Error (nm)
X	26	+/-17	60
Z	31	+/-11	53

Table 3: Overall results of positioning tests.

The increase in the consolidated Dead Zone Error occurred because of the difference in the locations of the fast positioning mean and the slow positioning mean. The effect of the traverse speed appeared to have a moderate impact on positioning along the X-axis: the Dead Zone for each X-axis position widened slightly while the Scatter Errors appeared to be the averages of the fast and slow traverse speeds. Figures 3 and 4 bear witness to the lesser impact of the speed change on X-axis positioning, showing that the Dead Zone for the grouped data is only slightly widened by the separation of the fast positioning mean and the slow positioning mean.

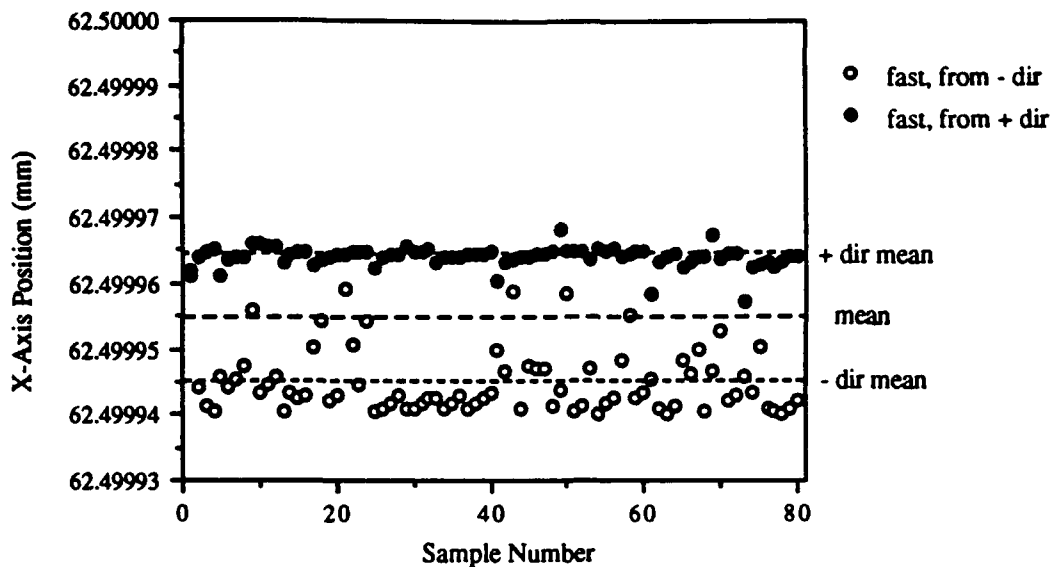


Figure 3: X-axis positioning (fast traverse) plot for nominal $X = 62.5$ mm.

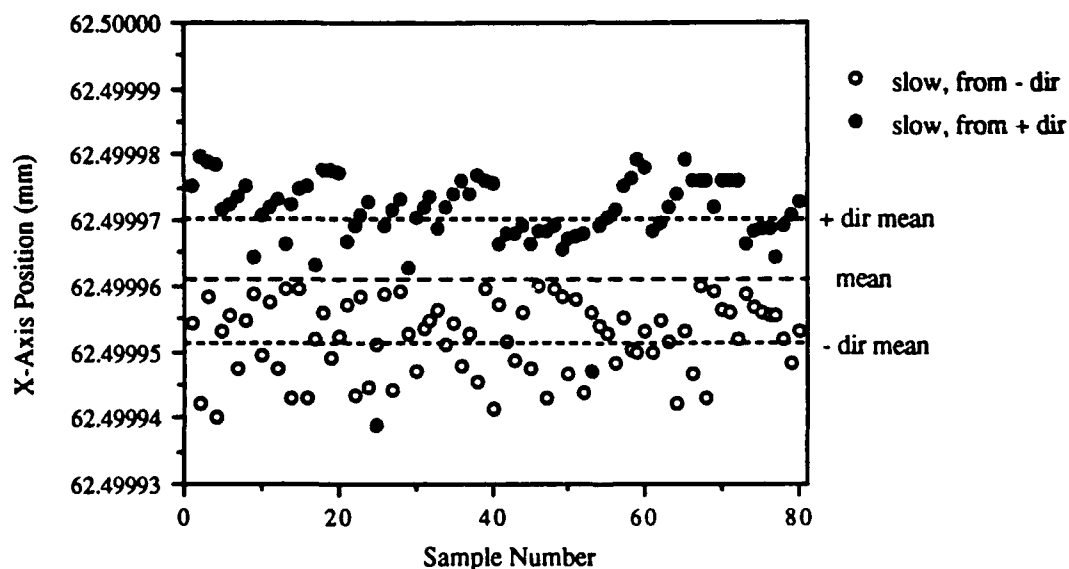


Figure 4: X-axis positioning (slow traverse) plot for nominal $X = 62.5$ mm.

The consolidated Scatter Errors for the Z-axis positions appeared to be the averages of the fast and slow traverse speeds. However, the consolidated Dead Zone Errors for the two Z-axis positions were significantly wider than any of the individual Dead Zones. The Z-axis Dead Zone Errors increased because in each case, the fast positioning mean was located 15-20 nm away from the slow positioning mean (see Figures 5 and 6). One explanation of the traverse speed's impact on the Z-axis but not the X-axis is that inertial effects are more pronounced for the Z-axis since the Z-

axis carriage/spindle mass is twice that of the X-axis carriage. Differences in the servo amplifier gains, biases, etc., could also be cited as possible causes.

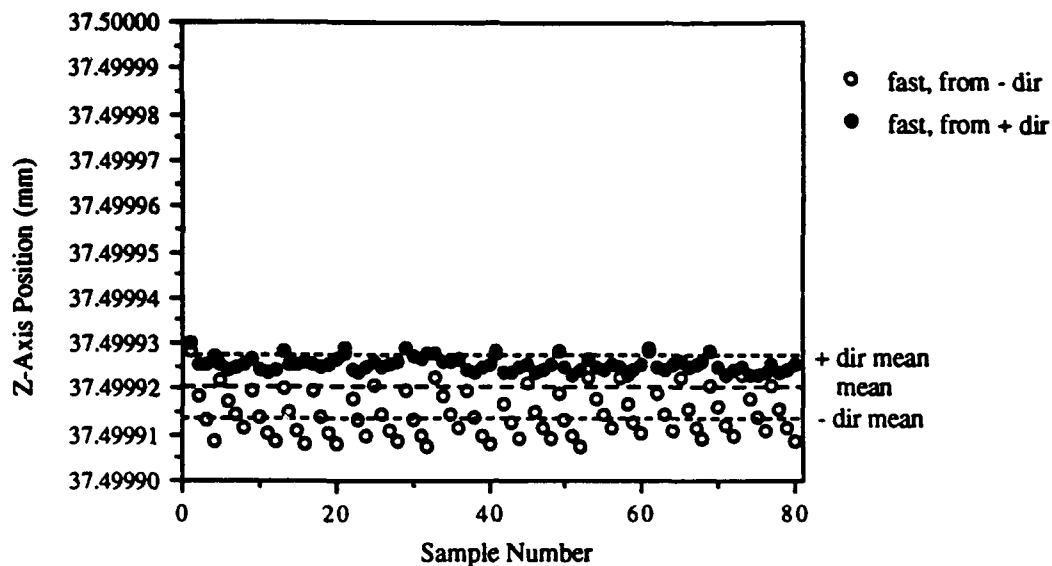


Figure 5: Z-axis positioning (fast traverse) plot for nominal $Z = 37.5$ mm.

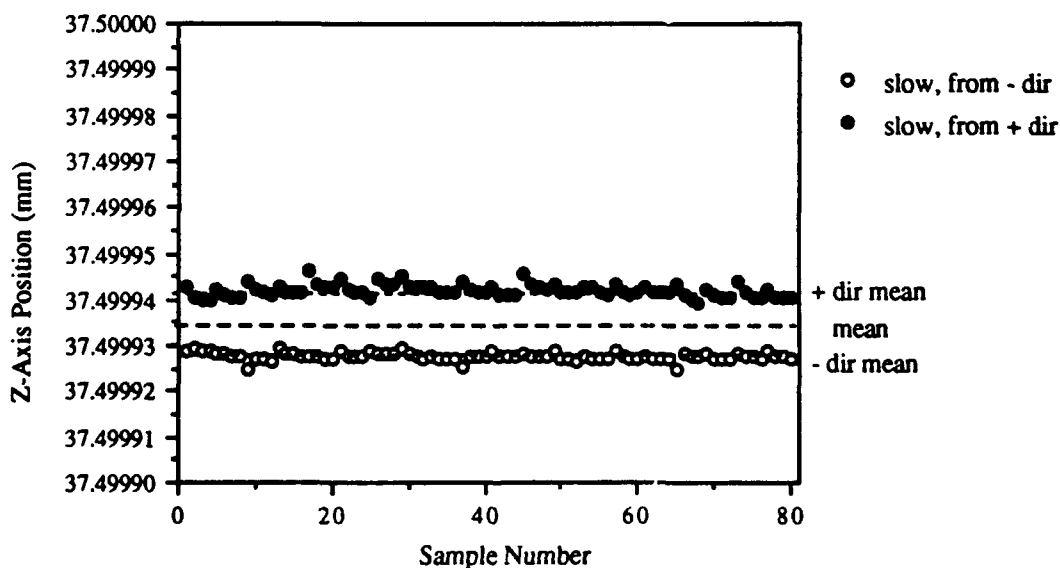


Figure 6: Z-axis positioning (slow traverse) plot for nominal $Z = 37.5$ mm.

Note that in Figures 3 through 6, the mean of the approaches from the positive direction and the mean of the approaches from the negative direction both fall on the negative side of the nominal position. Since the positioning control for this experiment used only proportional feedback, one might expect the "negative" mean to fall on the negative side of the nominal position and the

"positive" mean to fall on the positive side of the nominal position. However, because the servo motor amplifier is not completely balanced, the residual voltage bias will always result in a "drift" in a certain direction depending on the bias. Therefore, the actual axis position will consistently fall on the same side of the nominal position, independent of the direction of approach. The amount of the bias will also impact the magnitude of the Dead Zone.

The X- and Z-axis positioning uncertainties (60 nm and 53 nm, respectively) would limit the attainable surface finish to no better than one tenth of a wave¹, peak-to-valley. Note also that the magnitudes of these errors agree with Miller's 40 nm repeatability error in the axes straightness measurements (Section 16).

6.3 SPINDLE GROWTH TESTS

Thermal growth in bearings occurs when frictional losses produce heat. In the DTM's hydrostatic bearing spindle, air is sheared by the rotation of the journal relative to the bearing, thereby generating heat. The generated heat gives rise to thermal gradients which lead to differential expansion of the bearing components. The differential expansion is further complicated if the structure is composed of several materials with different coefficients of thermal expansion.

6.3.1 Setup and Procedure for Spindle Growth Test #1

Thermal growth of the DTM's air bearing spindle was measured using the setup shown in Figure 7. An ADE Microsense Capacitance Gage was used to measure the growth (or shrinkage) of the spindle as a function of spindle speed and time. The capacitance gage probe was mounted to the X-axis carriage and a diamond turned target was mounted to the vacuum chuck of the spindle. The DTM controller recorded the analog voltage signal from the capacitance gage, along with the actual carriage positions as measured by the DTM's laser interferometer system. (Note that during the data collection, the X- and Z-axes were not holding position, but rather both axes drive motors were disabled.) The output signal from the capacitance gage had a range of +/- 10 volts, with a scaling factor of 2.54 μm per volt. The capacitance gage and interferometer were sampled every millisecond; each minute of data (i.e., sixty thousand samples) was averaged to obtain one datapoint. The capacitance gage voltages were converted to position changes and plotted versus time.

With the spindle initially at rest and "cool", the capacitance gage signal was recorded for four hours. At the end of the first four hours, the spindle was brought to a rotational speed of 1000 rpm, clockwise. The capacitance gage signal was then recorded for two consecutive 12-hour

¹ As measured with Helium/Neon laser light of 632.8 nm wavelength.

intervals.² At the end of the second 12-hour interval, the spindle was turned off and the capacitance gage signal was recorded for three consecutive 16-hour intervals. The results of the test are shown in Figure 8. Note that an increase in the ordinate variable represents a decrease in the separation of the probe and the target (i.e., presumably representing growth of the spindle toward the cap gage probe.) Although the capacitance gage reading really represents the change in the separation rather than the actual separation, for simplicity the nominal probe stand-off distance (100-150 μm) will be ignored and the capacitance gage reading will be treated as the separation distance.

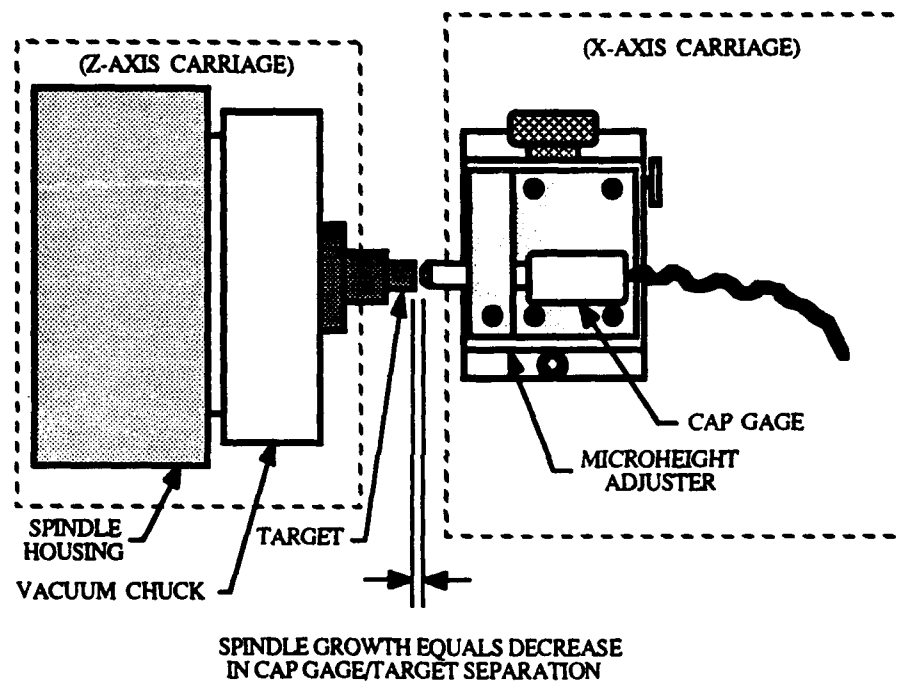


Figure 7: Equipment Setup for Spindle Growth Tests.

6.3.2 Results of Spindle Growth Test #1

In the first four hours of the test, there is a steady decrease in the separation distance which, given its 100 nm magnitude, could possibly be attributed to thermal growth due to the laboratory air temperature variations. An increase in the separation distance can be seen when the spindle was started at $t=4$ hours, possibly due to a shift in the axial position of the spindle journal or an increase in the convective heat transfer (due to the air movement associated with the rotating spindle).

² The sampling periods were divided into equal intervals because of constraints associated with the data collection routine.

However, after the initial increase in separation, the spindle appears to grow toward the probe until reaching a "steady-state" separation of 1150 nm at $t=12$ hours.

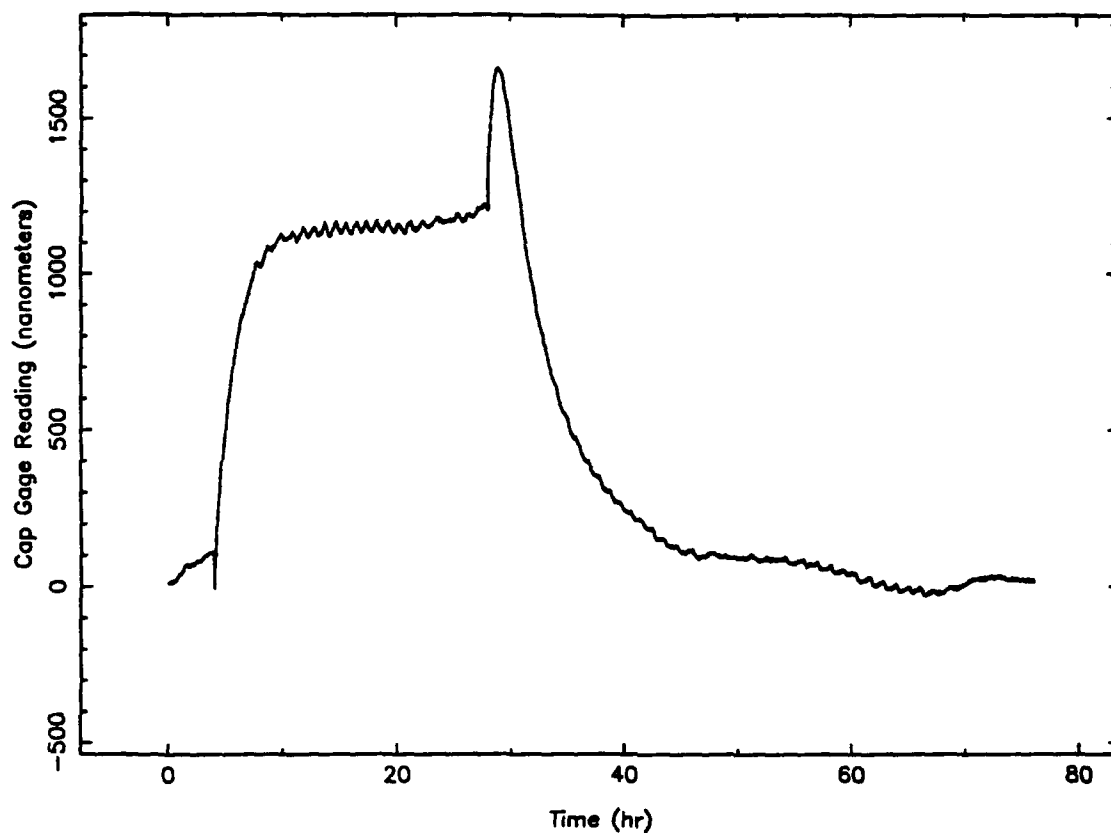


Figure 8: Spindle Growth Test #1 (76 hours in duration).

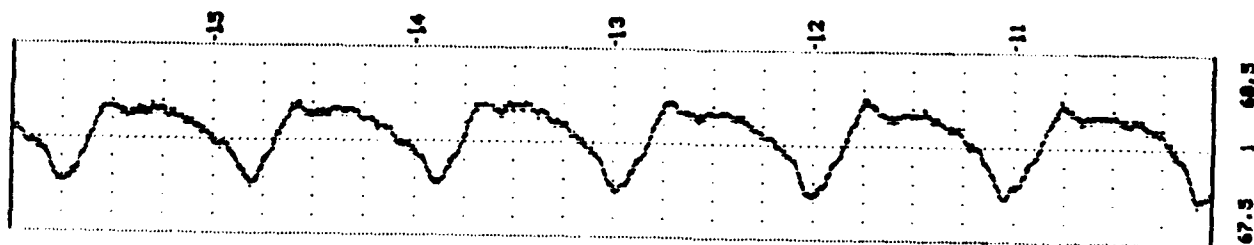


Figure 9: DTM Lab air temperature (°F) during hours 11 through 16 of Spindle Growth Test #1.

The hourly 0.3°C (0.5°F) variation in the laboratory air temperature³ shown in Figure 9 could possibly be the cause of the hourly 40-50 nm "steady-state" fluctuation shown in Figure 10. Substantiation of this hypothesis would require an analysis of the complex heat transfer mechanisms associated with the structural loop between the probe and target; however, the periods of the fluctuations agree, and given the effects of thermal inertia, the 72° phase lag between the fluctuations in the air temperature and the capacitance gage reading is not unreasonable.

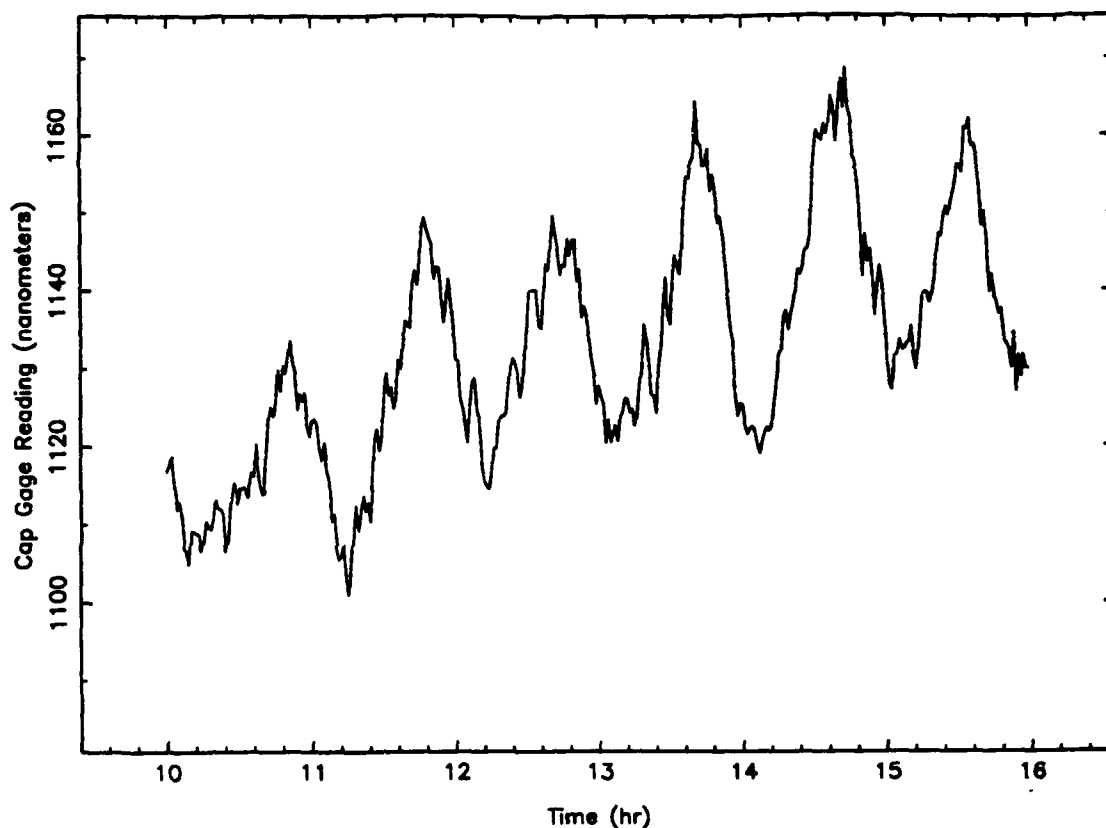


Figure 10: Spindle Growth Test #1 (hours 11 through 16).

When the spindle is stopped at $t=28$ hours, the separation begins to rapidly decrease another 450 nm over a one hour period. A fraction of this apparent additional growth could be attributed to another shift in the axial position of the spindle journal; however, because of the magnitude of the change, the predominant cause is more likely to be a decrease in the convective heat transfer. After one hour of cooling (at $t=29$ hours), the separation begins to increase until stabilizing (at $t=49$ hours) near the value of the original separation. The remaining 27 hours exhibit a fluctuation in the

³ An investigation into this temperature cycling revealed that the disturbances were caused by variations in the chilled water temperature in the system's cooling coil. Correction of this situation is documented in Section 6.4.

separation distance which again could be postulated to be the result of variations in the laboratory air temperature.

Based on this test's spindle growth measurement of nearly 1200 nm, operation of this DTM (at a spindle speed of 1000 rpm) should be preceded by a spindle warmup period of approximately eight (8) hours. Other spindle speeds may mandate longer or shorter warmup periods.

6.3.3 Setup and Procedure for Spindle Growth Test #2

Given the stopping and starting of the spindle associated with normal operation, a second test was performed to assess the effect of operation cycling on spindle growth. The 57-hour experiment used the same setup and sampling as the first test, and began with the spindle initially at rest and "cool". The capacitance gage signal was recorded for eight hours with no spindle rotation. At the end of the first eight hours, the spindle was brought to a rotational speed of 1000 rpm, clockwise, and the capacitance gage signal was recorded for twelve hours. Next, five 1-hour cycling tests (i.e., the spindle was turned off for one hour, then operated at 1000 rpm, clockwise, for one hour) were performed. At the end of the fifth 1-hour cycling test, the spindle speed was maintained at 1000 rpm, clockwise, for an additional three hours to allow the system to restabilize before beginning the five half-hour cycling tests. Each half-hour cycling test consisted of thirty minutes without spindle rotation followed by thirty minutes of 1000 rpm, clockwise spindle rotation. At the end of the fifth half-hour cycling test, the spindle speed was maintained at 1000 rpm, clockwise, for an additional three hours to allow the system to restabilize before stopping the spindle for the final sixteen-hour cool-down test.

6.3.4 Results of Spindle Growth Test #2

The results of the second spindle growth test are shown in Figure 11. In the first eight hours of the second test, there is a steady decrease in the separation distance up to 100 nm, possibly due to variations in the environmental conditions, etc. As seen in the first test, an increase in the separation distance can be seen when the spindle was started (at $t=8$ hours). However, after the initial increase in separation, the spindle appears to grow toward the probe until reaching a "steady-state" separation of 1500 nm after 10 hours of operation (i.e., at $t=18$ hours).

A closeup of the five 1-hour cycling tests is shown in Figure 12. After the spindle is stopped at $t=20$ hours, the separation begins to rapidly decrease (there is probably a heat buildup due to a decrease in the convective heat transfer away from the probe/target interface) another 350 nm over a period of 40 minutes. From that time, the separation increases slightly until the spindle is restarted at $t=21$ hours. When the spindle is restarted, the separation abruptly increases more than 500 nm over a 30-minute period. The separation remains nearly constant over the next 30 minutes until the spindle is stopped again. While the magnitudes of the changes in separation vary,

producing an overall trend toward a larger separation, the pattern is similar for all of the 1-hour cycling tests: the majority of the change (approximately 400 nm) occurs in 30 minutes followed by a small amount of drift.

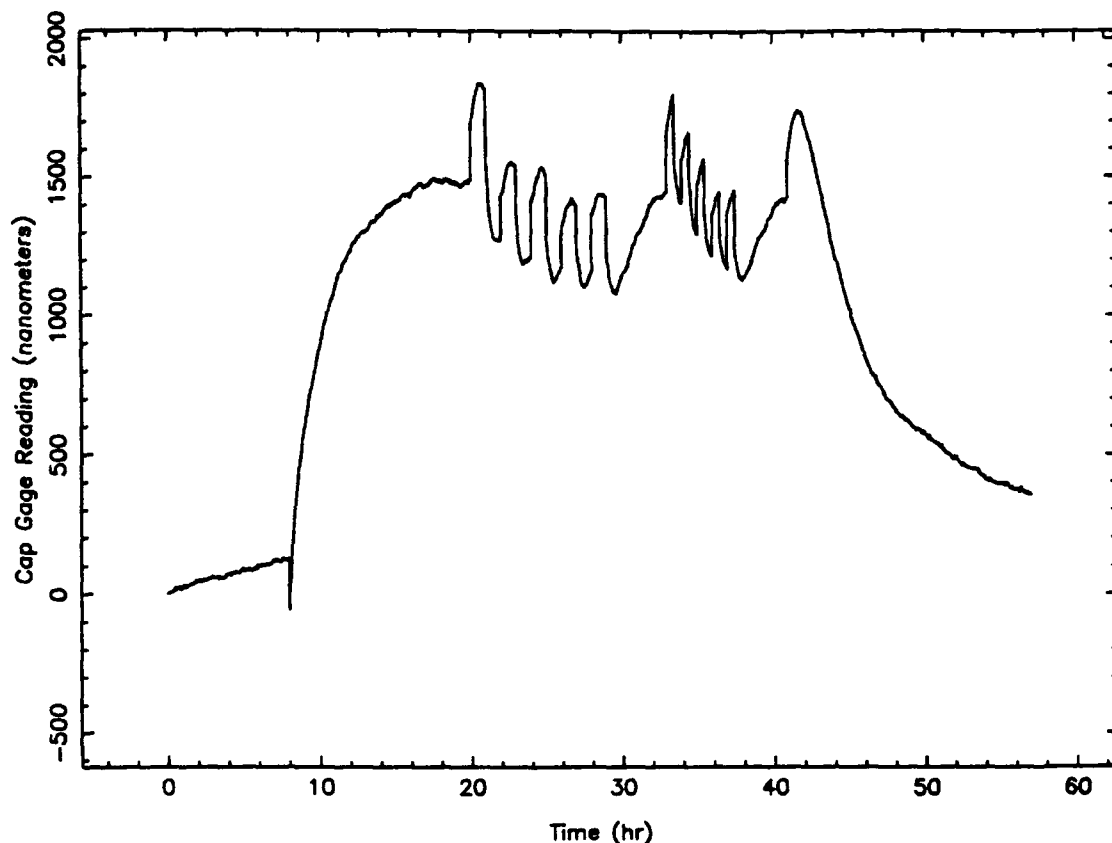


Figure 11: Spindle Growth Test #2 (57 hours in duration).

The results of the five half-hour cycling tests are also shown in Figure 12. During the 3 hours of operation allowed for restabilization, the separation reaches 1430 nm (nearly the "steady-state" separation of 1500 nm). After stopping the spindle for the first half-hour cycling test, the separation decreases 370 nm in the 30-minute "off" period. After operating the spindle for another 30 minutes, the separation returns to 1400 nm. The magnitudes of the separation changes for the five half-hour cycling tests range from 260 nm to 370 nm, resulting in an overall trend toward a larger separation. From these tests it appears that there is no great difference between stopping the spindle for 30 minutes and stopping the spindle for one hour.

To verify that the spindle growth value were not increased or decreased by drifting of the slides, the X-axis and Z-axis position curves were inspected. Figure 13 includes the X-axis and Z-axis position curves along with the spindle growth curve for the first 27.2 hours of the second test. Inspection of Figure 13 shows a large increase in the X-axis laser reading and a large decrease in the Z-axis laser reading over the 27.2-hour period.

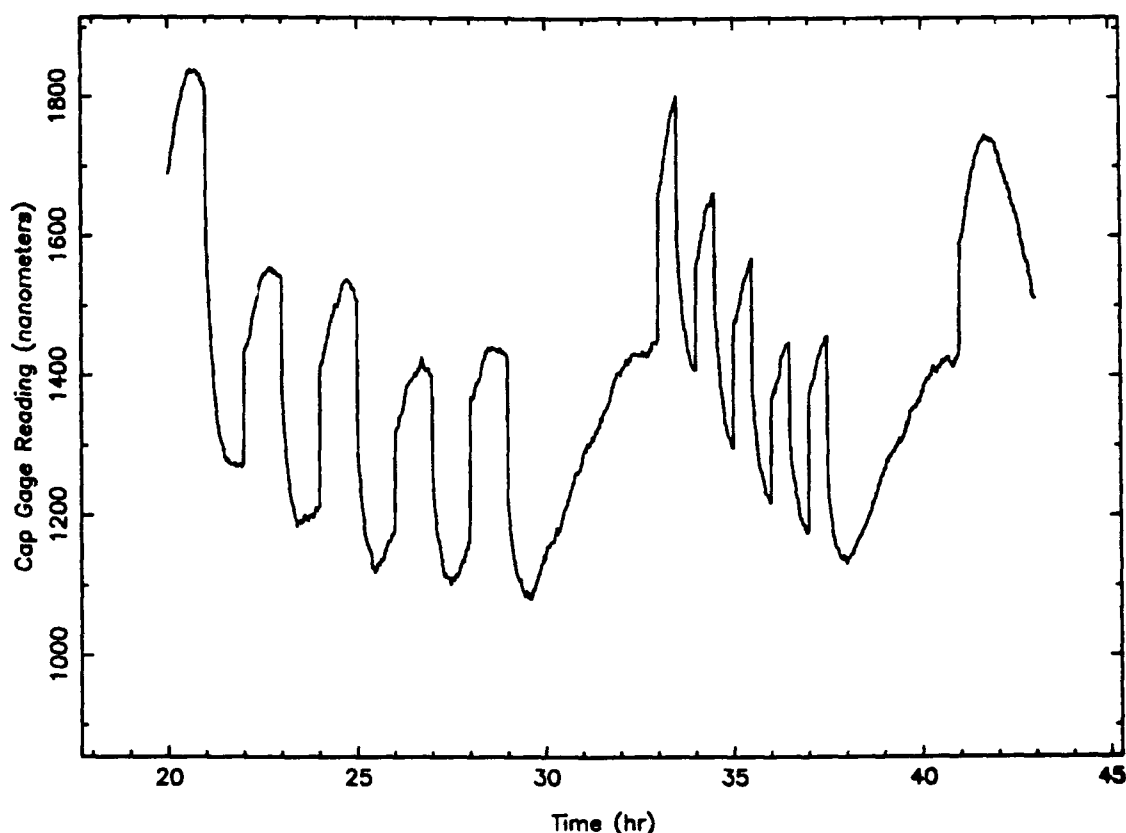


Figure 12: Spindle Growth Test #2 (hours 20 through 43).

6.3.5 Environmental Effects on Laser Interferometer

When the properties (e.g., temperature, barometric pressure, relative humidity, and carbon dioxide content) of an environment, or medium, change, the refractive index of the medium changes [2]. Since the refractive index is the ratio of the vacuum wavelength to the wavelength in the medium, the refractive index is indirectly proportional to the wavelength in the medium. The Zygo Axiom 2/20 two-frequency heterodyne laser interferometer (used on the Center's DTM) senses motion of the retroreflector relative to the interferometer by measuring the phase difference between the frequencies of the "optical interference measurement signal" and the electrical reference signal [3]. The optical interference measurement signal (or interference signal) is produced by the interference of the laser's reference and measurement beams, f_2 and f_1 , respectively. The frequency of the interference signal is given by $f_2 - f_1 \pm \Delta f_1$.

Movement of the retroreflector away from the interferometer results in a decrease in the frequency of the interference signal (because of the Doppler Shift in f_1) and a change in phase of the interference signal with respect to the electrical reference signal. Movement of the retroreflector

toward the interferometer results in an increase in the frequency of the interference signal and an opposite shift in the phase of the interference signal with respect to the electrical reference signal.

However, if the retroreflector is held stationary and the measurement path's refractive index decreases, the wavelength of the measurement beam will increase and the measurement beam (f_1) will experience a phase shift relative to the reference beam (f_2). This phase shift will change the phase of the interference signal relative to the electrical reference signal. Since the motion of the retroreflector is detected by measuring the relative phase of the interference and electrical reference signals, there will be an erroneous detection of retroreflector motion toward the interferometer. Conversely, if the measurement path's refractive index increases, the wavelength of the measurement beam will decrease (producing an opposite shift in the phase of the interference signal relative to the electrical reference signal), thereby resulting in the false detection of retroreflector motion away from the interferometer.

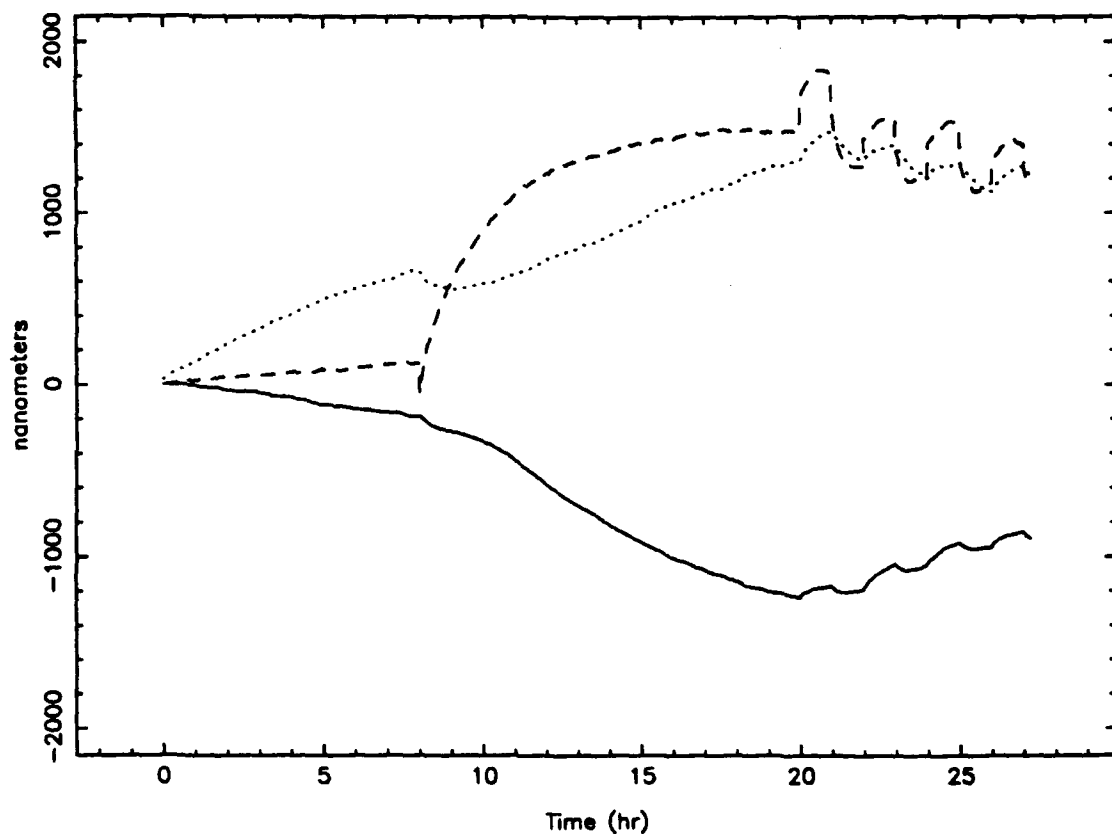


Figure 13: Spindle Growth Test #2 (first 27.2 hours). The X-axis laser interferometer position values are represented by the dotted line, the Z-axis laser interferometer position values are represented by the solid line, and the spindle growth values are represented by the dashed line.

To determine whether the laser readings changed because of slide motion or uncompensated refractive index changes, the environmental conditions for $t=0$ and $t=27.2$ hours were examined. Table 4 lists the temperature, relative humidity, and barometric pressure data measured for the two times. The refractive indexes which would result from the two sets of environmental conditions were calculated using the Edlen Equation [2]:

$$n_{\text{air}} = 1 + \frac{(0.27651756 \times 10^{-3})(1 + 0.54 \times 10^{-6}(C - 300))(0.104127 \times 10^{-4}(P))}{(1 + 0.3671 \times 10^{-2}(T))} - 0.42063 \times 10^{-9}(F) \quad (9)$$

where C is the carbon dioxide content in parts per million, F is the water vapor pressure in Pascals, P is the air pressure in Pascals, and T is the air temperature in °C. Then the X-axis and Z-axis laser reading errors were estimated using the two calculated refractive indexes and the X-axis and Z-axis cavity lengths (i.e., the distance between the interferometer and the moveable retroreflector).

Time of Measurement	DTM Enclosure Air Temperature	Lab Air Vapor Pressure	Lab Air Pressure	Estimated Carbon Dioxide Content [2]
$t = 0$	21.51°C	1069 Pa (RH=42.0%)	102014 Pa (BP=766.9 mm Hg)	750 ppm
$t = 27.2$ hrs	22.04°C	1379 Pa (RH=51.3%)	100056 Pa (BP=752.1 mm Hg)	750 ppm

Table 4: Environmental Changes for Spindle Growth Test #2.

The calculated laser interferometer errors (associated with the changes in environmental conditions from Table 4) were -1060 nm for the X-axis and +670 nm for the Z-axis. The recorded position changes for that time interval were +1170 nm for the X-axis and -900 nm for the Z-axis. Because of the disagreement of the signs of the calculated and measured values, the changes in the axes positions have been attributed to environmentally-induced laser interferometer errors *and* some unidentified counteracting process(es). For example, thermal expansion of the machine slides and base may have been partially responsible for the axes position errors.

Since these calculations have used laboratory air and DTM enclosure air conditions to approximate the actual laser air path conditions, the exact error values may differ. Furthermore, all of the axes position error cannot be attributed to environmentally-induced laser interferometer error. Nevertheless, these approximations do indicate that there is measurable environmentally-induced laser interferometer error and that a refractometer is necessary to compensate for the environmental effects.

6.4 LABORATORY AIR TEMPERATURE CONTROL

The four laboratory HVAC systems at the PEC were designed to maintain the laboratory air temperatures at 20°C to within $\pm 0.06^\circ\text{C}$. Because the heat gains from the laboratories' occupants and equipment were unknown (and variable), the system was designed to cool the return and makeup air down to a prescribed temperature, then reheat the air to obtain the necessary supply air temperature as dictated by the laboratory air temperature thermistor.

Each laboratory is equipped with its own air handler, chilled-water-to-air heat exchanger, hot-water-to-air heat exchanger, PID control loops, and sensors (see Figure 14 and Table 5). Chilled water is delivered to the four local loops and cooling coils via an intermediate recirculating supply loop, which is in turn supplied by the primary recirculating chiller/storage tank loop. Hot water is delivered to the four local loops and heating coils via a primary recirculating supply loop which is heated by a steam-to-water heat exchanger.

Sensor #	Setpoint	Sensor Type	Location
S ₁	6.67°C (+1.39°C, -0.83°C)	Resistive Temperature Device (Time Constant: 5 sec for water @ 3 fps)	Primary Chilled Water Supply Loop on output side of chiller
S ₂	10.6°C	Resistive Temperature Device (Time Constant: 5 sec for water @ 3 fps)	Intermediate Chilled Water Supply Loop on output side of circulating pump
S ₃	18.6°C (Typ.)	Thermistor (Time Constant: 20 sec for air @ 20 fps)	Air Duct on output side of cooling coil
S ₄	20°C	Thermistor (Time Constant: 20 sec for air @ 20 fps)	Laboratory
S ₅	30.6°C	Resistive Temperature Device (Time Constant: 5 sec for water @ 3 fps)	Primary Hot Water Supply Loop on output side of steam-to-water heat exchanger
S ₆	Not Used (Typ.)	Thermistor (Time Constant: 20 sec for air @ 20 fps)	Air Duct on output side of reheat coil

Table 5: Original Laboratory HVAC System Setpoints.

Originally, the laboratory HVAC systems failed to meet the temperature specifications. As detailed in Section 6.3.2 and shown in Figure 15, the laboratory air temperature would experience a periodic fluctuation. These variations could be characterized as hourly 0.3°C fluctuations, although the period and magnitude of the disturbances were variable (in fact, there were occasions when the disturbances would disappear completely).

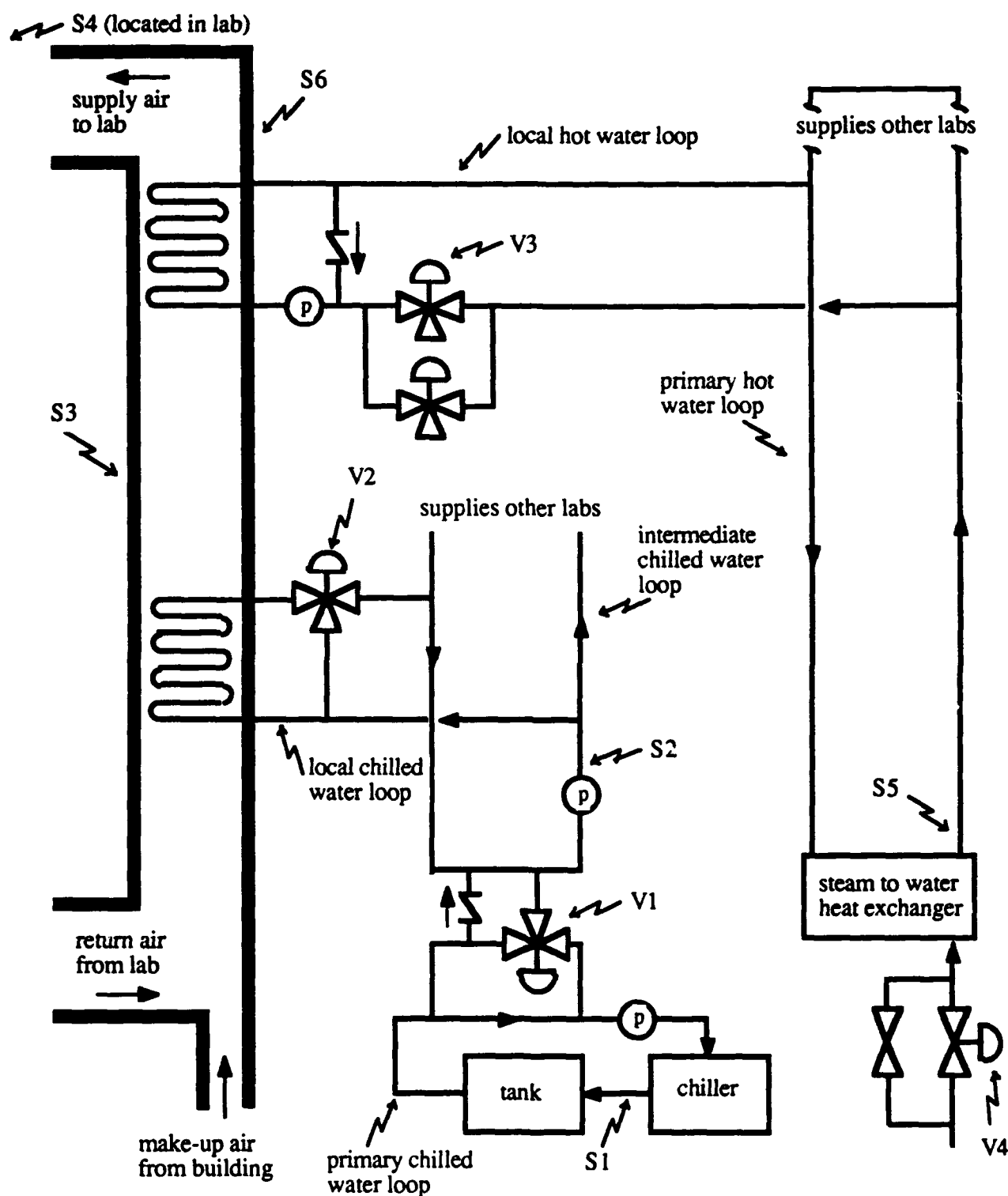


Figure 14: Single Laboratory HVAC System Schematic. (*P* denotes circulating pump, *S* denotes temperature sensor, and *V* denotes pneumatically-controlled valve.)

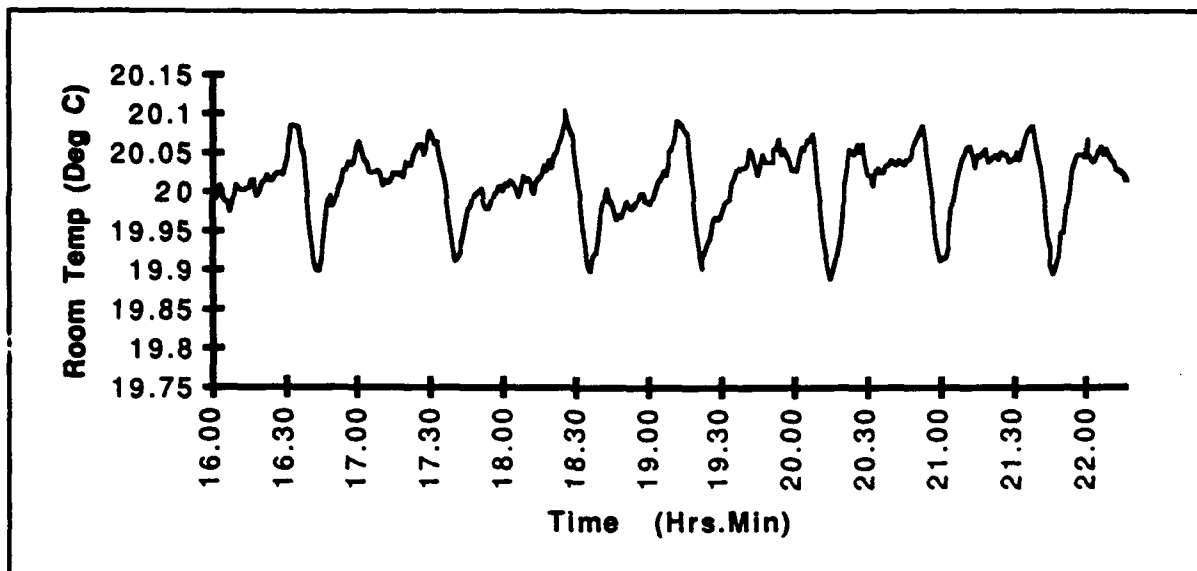


Figure 15: Laboratory Air Temperature Fluctuations.

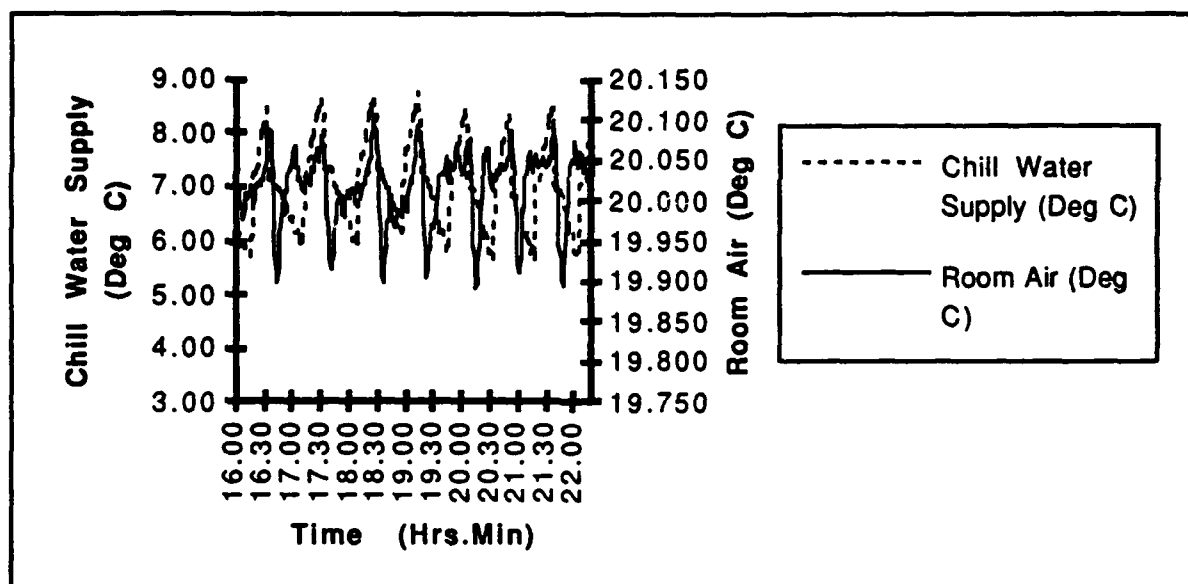


Figure 16: Laboratory Air Temperature and Primary Supply Loop Chilled Water Temperature Fluctuations.

The source of the disturbances was traced to variations in the temperature of the primary chilled water supply loop. In Figure 16, each peak in the Room Air Temperature is preceded by a peak in the Chilled Water Supply Temperature. The temperature cycling in the Primary Chilled Water Supply Loop was determined to be the result of the loading and unloading of the 3-stage compressor. The building HVAC control system has only on/off control of the chiller; the chiller's microprocessor controller independently loads and unloads the various compressor stages as it tries

to maintain the 6.67°C setpoint temperature. The temperature fluctuations were attributed to the decision deadband error associated with the loading and unloading of each stage.

With the original controller gains and setpoints, there was insufficient cooling capacity at the cooling coils when the chilled water temperature in the primary supply loop exceeded 7.8°C. That is, when the chilled water temperature (S_1) in the primary supply loop exceeded 7.8°C, even the fully open valve V_1 could not supply enough chilled water to maintain the intermediate loop setpoint (S_2). In turn, the elevated chilled water temperature in the intermediate loop would prevent the local loop from providing enough chilled water to the cooling coil to maintain the cooled air at its setpoint (S_3).

Several remedies for this problem were identified and investigated. First, the setpoint for the primary chilled water supply loop was lowered so that the chilled water temperature would never exceed 7.8°C. However, when the setpoint was dropped from 6.67°C to 5.56°C, the water temperature fell to 3.89°C, triggering a freeze alarm. This situation was obviously unacceptable so the setpoint was returned to 6.67°C.

The second option was to adjust the chiller's " ΔT " setting to minimize the decision deadband. As previously mentioned, the chiller operates at three discrete load levels; by decreasing the " ΔT " setting, the higher load levels are invoked sooner to return the temperature to the setpoint value more rapidly. This effectively adjusts the response of the chiller. When the " ΔT " setting was lowered to 3.33°C from 4.44°C, the water temperature was still seen to exceed 7.8°C. (Although this option did not alleviate the problem, the " ΔT " setting was left at 3.33°C.)

The third alternative would be to increase the storage capacity (presently 350 gallons) in the primary loop, thereby matching the time constant of the primary loop with the present response of the chiller. For budgetary reasons this alternative was designated the last resort and was not pursued.

The fourth option was to improve the disturbance rejection capability of each laboratory control system by adjusting the gains and setpoints on the cooling and heating coils. Through monitoring of the intermediate chilled water temperature (S_2), the cooling coil control valve pressure (V_2), and the "post cooling coil" air temperature (S_3), it became apparent that the cooling control valve was not opening to full capacity when the intermediate chilled water temperature rose above the 10.6°C setpoint. That is, the cooling coil had excess flow capacity which could be used to counteract the rising water temperature, but the control valve was not opening fast enough. The cooling coil control was retuned to obtain a faster response, and thereby reduced the system's sensitivity to the disturbances in the chilled water supply temperature. However, at full flow capacity, the cooling coil was still unable to consistently maintain the "post cooling coil" air temperature setpoint (S_3).

Therefore, the "post cooling coil" air temperature setpoint (S_3) was raised to 18.9°C so that the operation of the cooling coil control valve would not be near the valve saturation limit.

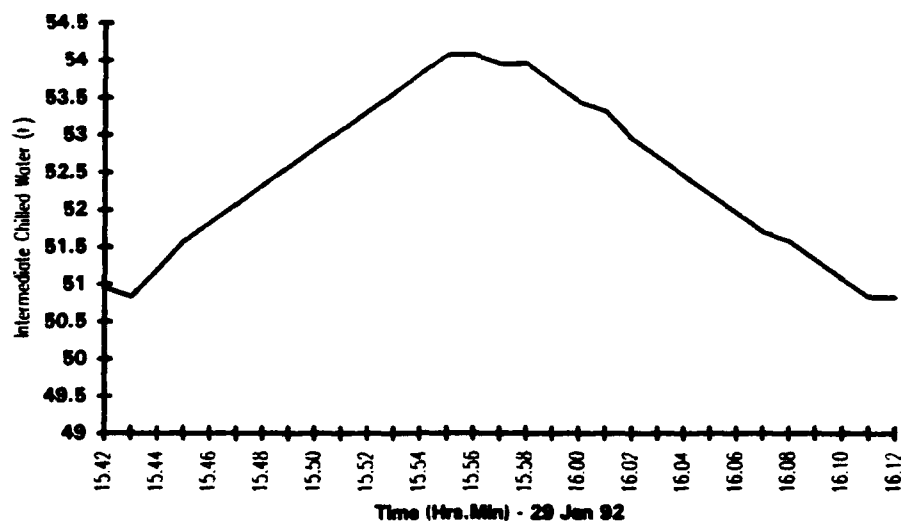


Figure 17: Chilled Water Temperature during the simulated disturbance in the Intermediate Chilled Water Loop.

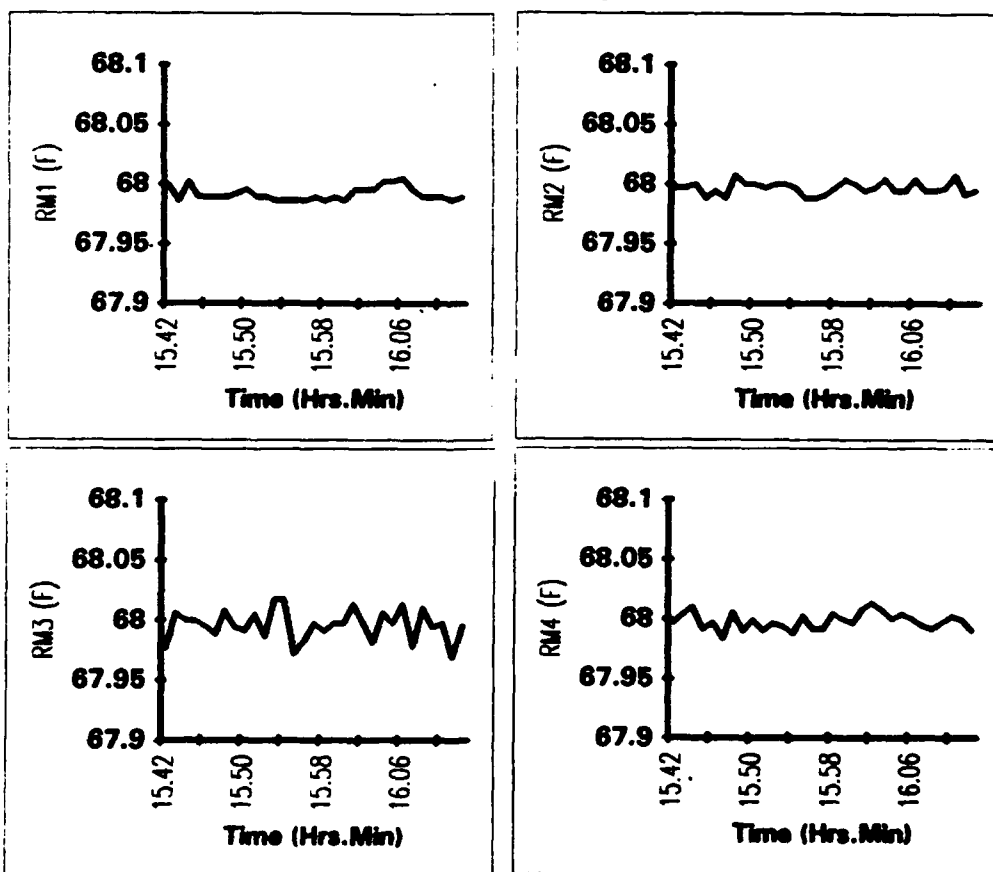


Figure 18: Laboratory Air Temperatures during the simulated disturbance in the Intermediate Chilled Water Loop.

To further improve the system's disturbance rejection capability, the control of the reheat coil was examined. The investigation showed that the relatively fast response of the reheat coil was not being maximized because the distant room air temperature sensor (S₄) was being used as the feedback variable instead of the nearby "post reheat coil" air temperature sensor (S₆). By reconfiguring the feedback loop (one laboratory already used the preferred configuration) and adjusting the gains, the disturbance rejection capability of the reheat loop was increased.

To assess the performance of the modified control system, a disturbance in the chilled water supply temperature was simulated by ramping the setpoint of the intermediate chilled water loop from 10.6°C to 12.2°C at 0.14°C per minute (see Figure 17). After maintaining 12.2°C for approximately four minutes, the water temperature was ramped back down to 10.6°C. Figure 18 shows that despite the disturbance in the intermediate chilled water loop temperature, all laboratory air temperatures were held to within the specifications.

6.5 TOOL CENTERING

Texas Instruments described a problem encountered when using interferometric analysis to center a diamond tool in a DTM. Using an iterative technique, the tool position is adjusted until the interferogram of the workpiece reveals a nearly perfect sphere[4]. The error indicated by the interferogram is attributed to tool centering error.⁴

tool rake angle:	0°
tool radii:	0.020" & 0.030"
centerplug diameter:	0.532"
radius of curvature:	0.75"
angle of tool sweep:	20.77°
measurement aperture (w/ 1.5 f#):	0.5"
measurement aperture (w/ 3.3 f#):	0.227"

Table 6: Pertinent information on tool centering investigation by Texas Instruments and PEC.

The problem for Texas Instruments involves using a material for the centering procedure other than the actual material to be machined. Initially, a 0.030" radius diamond tool is centered by cutting a spherical surface on an aluminum centerplug. When the aluminum centerplug is replaced with a

⁴ However, it is important to note that, in reality, the machine positioning errors and the tool geometry errors will contribute to the measured error.

germanium centerplug, the resulting interferogram indicates that the tool is off-center by as much as 100 $\mu\text{in.}$ (2.5 μm). When the aluminum centerplug is recut, the tool once again appears to be centered. However, this phenomena does not occur to the same degree while machining with a 0.020" radius diamond tool. Based on these conditions, the issue appears to be a function of workpiece material and tool geometry.

	0.020" Radius Tool		0.030" Radius Tool	
	Convex	Concave	Convex	Concave
Centering Error	0 (centered)	0 (centered)	+0.0001" (short)	-0.0001" (past)

Table 7: Texas Instruments' results for cutting a germanium centerplug after centering the tool with an aluminum centerplug.⁵

Tool cutting forces are dependent on both workpiece material and tool geometry. An early hypothesis to explain the centering change was that differences in workpiece stiffness combined with differences in cutting force were producing differences in deflections of the tool and centerplug. Tool forces for cutting germanium and aluminum were measured using the PEC's tool force dynamometer [5]. The results of these tests are shown in Table 8. Based on the force data, the calculated differences in tool and centerplug deflections do not appear to be of the magnitude required to produce the centering errors.

Cutting Conditions:	Tool Radius:	Cutting Aluminum:	Cutting Germanium:
DOC=0.00005" Feed=0.05 ipm N=1500 rpm	0.030"	$F_y=0.007\text{N}$ $F_z=0.014\text{N}$	$F_y=0.015\text{N}$ $F_z=0.090\text{N}$
	0.020"	$F_y=0.008\text{N}$ $F_z=0.010\text{N}$	$F_y=0.014\text{N}$ $F_z=0.080\text{N}$
DOC=0.0001" Feed=0.2 ipm N=1500 rpm	0.030"	$F_y=0.032\text{N}$ $F_z=0.042\text{N}$	Not Measured
	0.020"	$F_y=0.033\text{N}$ $F_z=0.028\text{N}$	Not Measured

Table 8: Tool force data (measured 05/21/91 and 05/22/91 at PEC).

Cutting tests will be performed on the materials listed in Table 9 to determine the significance, if any, of elastic modulus and hardness in the centering error phenomena. The materials and their

⁵ As measured with a Mark II Laser Interferometer using a 1.5 F-number transmission sphere.

heat treatments (if applicable) were selected to provide a range of elastic moduli and a range of hardnesses. The selections were made such that the order of the materials in the elastic modulus range would not correspond to the order of the materials in the hardness range.

Workpiece Material:	Elastic Modulus:	Hardness:
6061-T6 Aluminum	69 GPa [6]	Brinell 95 [6]
Annealed OFHC Copper	117 GPa [6]	Rockwell F 40-45 [6]
Germanium	128 GPa [7]	Vickers 900 [8]
Electroless Nickel	Not Available	Not Available

Table 9: Workpiece Material Properties.

The tests will be performed as shown in Table 10. Force measurements will be recorded during the cutting processes and the resulting centering errors will be measured using the Zygo Mark IV Laser Interferometer and the Zygo Maxim 3D Interferometric Microscope. This information will then be analyzed to determine any correlation with elastic modulus or hardness.

Cutting Conditions:	Spherical Radius:	Tool Radius:	Workpiece Material:
DOC=0.00005" Feed=0.05 ipm N=1500 rpm	0.75" Convex	0.030"	Aluminum
			Germanium
			OFHC Copper
			Electroless Nickel
		0.020"	Aluminum
			Germanium
			OFHC Copper
			Electroless Nickel
DOC=0.0001" Feed=0.2 ipm N=1500 rpm	0.75" Convex	0.030"	Aluminum
			Germanium
			OFHC Copper
			Electroless Nickel
		0.020"	Aluminum
			Germanium
			OFHC Copper
			Electroless Nickel

Table 10: Proposed Test Conditions.

References

- [1] Tlustý, Jiri, "Testing of Accuracy of NC Machine Tools", *Supplement 1 of Technology of Machine Tools*, October, 1980.
- [2] Garrard, K.P., Taylor, L.W., Knight, B.F., Fornaro, R.J., "Diamond Turning Machine Controller Implementation", *Precision Engineering Annual Report*, Volume 6, 1988.
- [3] "Axiom 2/20 Laser Measurement System Operation and Reference Manual OMP-0220", Zygo Corporation, April 1988.
- [4] Gerchman, M. C., "Optical Tolerancing for Diamond Turning Ogive Error", *Reflective Optics II*, Proceedings of SPIE Conference, 1989, Volume 1113.
- [5] Drescher, J. D., "Measurement of Tool Forces in Diamond Turning", *Precision Engineering Annual Report*, Volume 6, 1988.
- [6] "Materials Selector 1989", *Materials Engineering* (December, 1988).
- [7] Hirth, J. P. and Loethe, J., "Theory of Dislocations", J. Wiley, NY (1982).
- [8] Davidson, D. L. and Lankford, J., "The Crack Initiation Threshold in Ceramic Materials Subject to Elastic/Plastic Indentation", *Journal of Material Science*, 1979, Volume 14.

7 TOOL FORCE, SURFACE FINISH ASPECTS IN DIAMOND TURNING OF DUCTILE METALS

Joseph D. Drescher

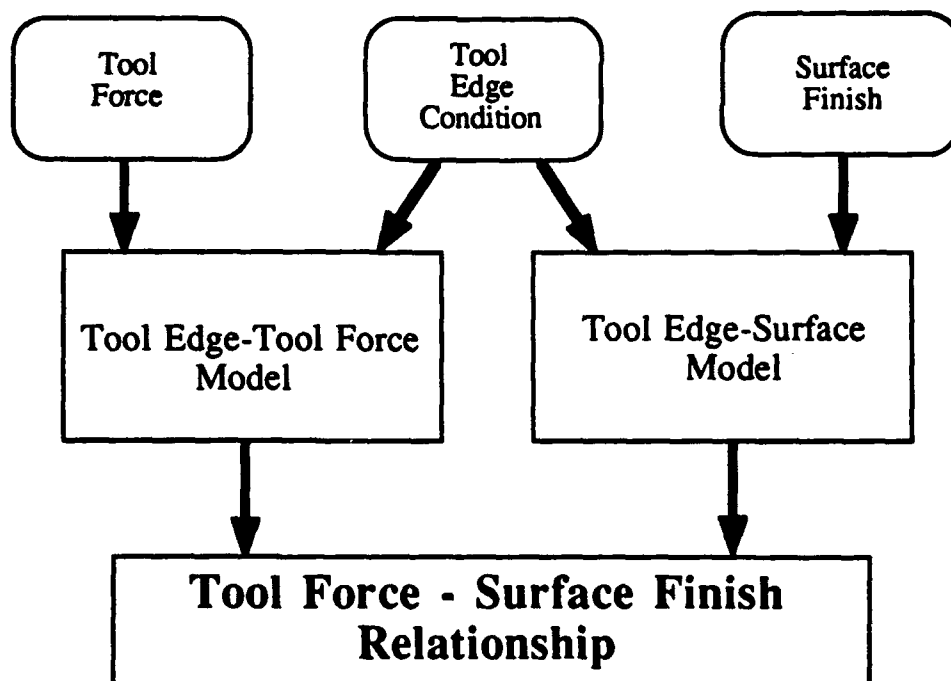
Graduate Student

Thomas A. Dow

Professor

Mechanical & Aerospace Engineering Department

The research described in this section involves tool force and surface finish aspects of diamond turning. It is shown that these two are related to one another since the tool edge condition affects both force and finish. Tool forces are described in terms of tool wear and the material properties, strength and ductility. It is also shown that tool wear can be related to surface finish in fine machining with a round nose tool by the concept of a minimum depth of cut which depends on tool edge radius, increasing with wear.



7.1 INTRODUCTION

7.1.1 Previous work in tool force modeling

A model of tool forces including material properties, tool geometry and cutting parameters is being developed. The goal is to be complete and general in this development by relying on empirical relationships as little as possible. Carroll [1] showed that the force on the tool in the cutting direction was related directly to the rate of material removal for a given cutting speed. He also found that this "cutting" force was proportional to the workpiece material's yield strength.

There have been two main advances in the improvement and refinement of Carroll's model for tool forces. The first was to account for forces both in the cutting direction and in a direction normal to this, in the "thrust" direction [2]. The thrust force tends to push the tool away from the workpiece. The second improvement was to recognize the effects of tool wear which contribute to measured force magnitudes in both directions [3]. When depth of cut is small relative to edge sharpness, measured forces do not scale linearly with depth of cut. This is due to contact on the tool edge behind the rake face caused by tool wear and elastic springback of the workpiece material. Similar to the normal and tangential forces acting on the rake face of the diamond, there is a pair of force components acting on this area of contact at the tool surface between the rake and clearance faces. Therefore, the measured force in the cutting direction is a combination of normal force on the rake face and frictional force on the wear land. The measured force in the thrust direction is normal force on the wear land plus friction on the tool rake face. Forces due to wear are not related to depth of cut. This model of force components has been related quantitatively to measurements of the tool edge.

7.1.2 Current research emphasis

A major emphasis of this research project is to understand how all cutting variables affect tool force. The variables of interest are not limited to tool parameters such as nose radius, rake angle, or clearance angle nor to machine parameters such as depth of cut, cutting velocity, and feedrate. It is important to understand the influence of these tool and machine parameters but the model must also include tool wear and properties of the work material in a quantitative way. These latter variables seem to be most important to the quality of the finished surface.

The second major emphasis is to establish the factors which determine surface finish and how they are related to tool forces. It is well known that the theoretical surface, as calculated assuming a perfect tool, ideal material, and perfect machine motion with no vibration, can not be achieved in diamond turning. It is necessary to see how the tool edge condition, i.e. the extent of tool wear, affects the finish. The tool condition is the key link between tool force and surface finish.

7.2 MODELING OF TOOL FORCE IN ORTHOGONAL CUTTING GEOMETRY

7.2.1 Force Measurement System

Figure 1 is a schematic of the force measurement system. The main component is a piezoelectric, quartz transducer which is sensitive in three perpendicular directions. A preload is required so that the transverse forces can be transmitted by friction. This preload is applied by a 6 mm bolt through the center of the transducer clamping it between the toolholder and dovetailed block. Time varying temperature changes cause strains of the piezo-electric element in all directions which can not be separated from strains caused by cutting force. This effect is greatest along the preload axis. In the cutting force measurements described in this report, forces were measured in the two directions perpendicular to the preload axis. In this way the temperature effects on force were reduced. The directions of interest are labeled Y and Z in Figure 1.

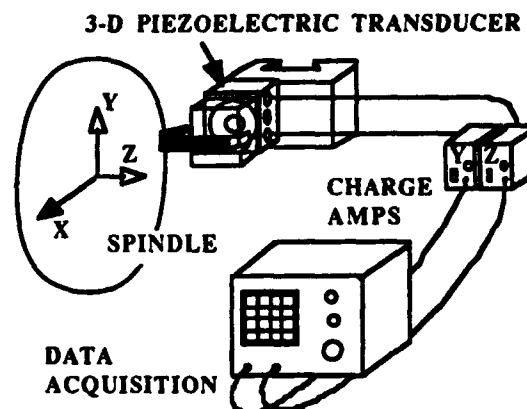


Figure 1: Force measurement system schematic

7.2.2 Tool force model

The schematic in Figure 2 shows the orthogonal cutting geometry and the tool force components considered in the model. The chip produced has uniform thickness and width. In this orthogonal geometry, the cutting parameters controlled by the machine are limited to cutting depth and edge length. The feed is in the Z-direction so that feed/revolution is equal to cutting depth, d . The edge length, L , is determined by the length of tool employed in the cut. 0° rake and 6° clearance angle tools were used in the work reported here.

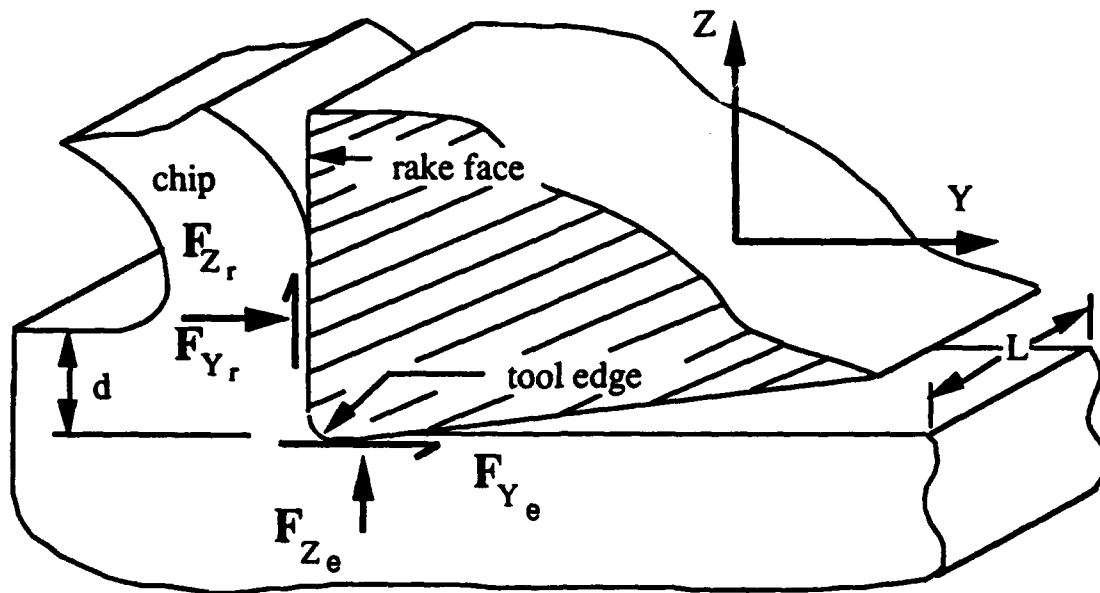


Figure 2: Orthogonal tool force model

From Figure 2 the following two equations can be written to describe the measured tool forces.

$$F_Y = F_{Yr} + F_{Ye} = C_1 d L + \mu C_2 L \quad (1)$$

$$F_Z = F_{Zr} + F_{Ze} = \mu C_1 d L + C_2 L \quad (2)$$

Equations (1) and (2) have three unknown variables, C_1 , C_2 , and μ . C_1 relates normal force on the rake face to the cross-sectional area of the uncut chip. C_2 relates the normal force on the edge of the tool to the edge length in contact with the workpiece. An increase in C_2 corresponds to either an increased stress on the clearance face wear land or an increase in the wear land area. From the units it can be seen that C_2 is a stress multiplied by length. The friction coefficient, μ , on the rake face is assumed to be equal to μ on the clearance face.

This force model, Equations (1) and (2), is modified slightly from that reported previously [3]. A fourth parameter was included in the earlier work as an exponent on the depth of cut, d . This was used to account for nonlinearities in force with depth of cut in the preliminary data. However, additional data has shown that this exponent is very close to unity and is unnecessary. Therefore the model was simplified.

7.2.3 Correlation with experiment

With experimental force data, where d , L , F_Y , and F_Z are known, C_1 , C_2 , and μ are found which optimize the correlation between measured forces and the model. Figure 3 shows measured tool force data (points) in the Y-and Z- directions as a function of the depth of cut. The best fit lines are also plotted using Equations (1) and (2). The values for the friction coefficient ($\mu=0.21$), and the parameters (C_1 & C_2) are included in the graph as well as the correlation coefficient of the data fit. A correlation coefficient greater than 0.95 represents a good data fit. If the optimization program were run using any three of the force measurements, represented by the data points, the correlation would be 1.0. Only three equations are needed to get values for the three unknowns. However, the high correlation obtained using all force data means that any three data points would yield essentially the same values for C_1 , C_2 , and μ .

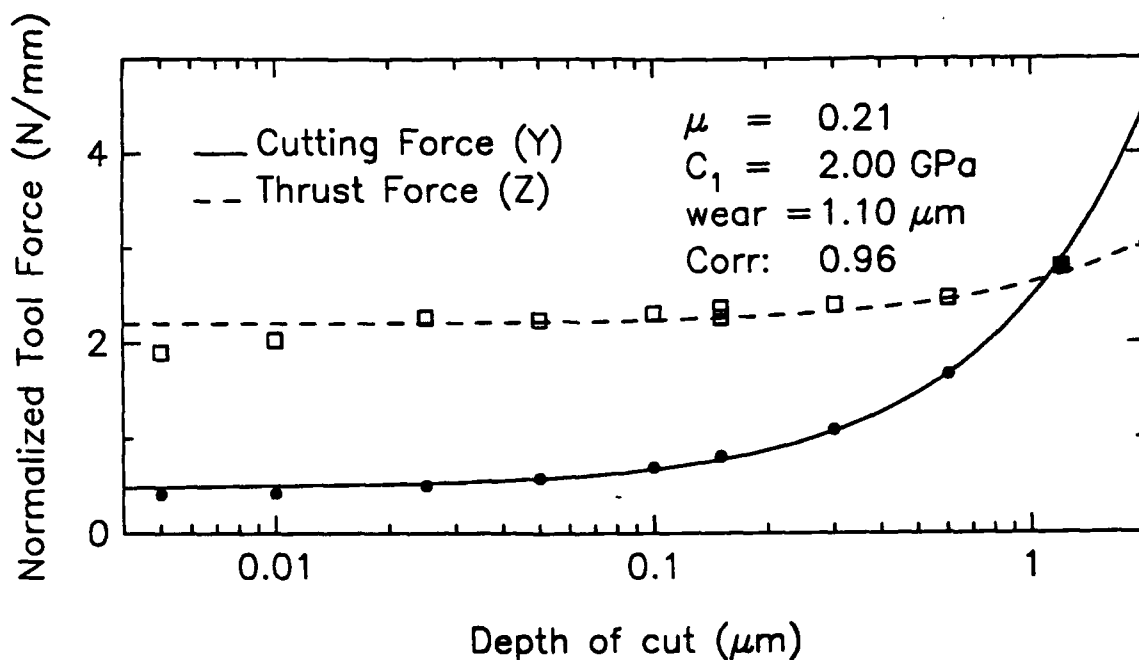


Figure 3: Experimental cutting force data (points) vs depth of cut plotted along with forces calculated using the best fit parameters in Equations (1) and (2).

The goal in modeling is to relate theory to experiment. Relying on empirical relationships is sometimes necessary but it limits the model's generality. In the force model, the physical significance of the friction coefficient is readily apparent. Also, the units on C_1 and C_2 give insight into their physical significance. C_1 is the stress acting on the rake face area, $L d$. C_2 (force / unit width) when multiplied by edge length, L , gives the normal force on the tool edge in the thrust direction. The width, w , included in C_2 is therefore perpendicular to the edge. The force $C_2 L$ represents a stress acting on the area, $L w$. The stress on this area is therefore C_2 / w . Both C_1

and C_2 should be related to material properties of the workpiece. C_2 should also be related to tool wear while C_1 should not change with wear.

7.2.4 Tool wear effects

Data similar to Figure 3 was obtained for several tools at various stages of tool wear. The C_2 required to fit force data increases as the tool wears while the other two parameters do not. This is illustrated in Figure 4, reproduced from a previous report [2]. Additional testing has been performed in which the tool edge was measured at the various stages of wear. By noting how C_2 changed as the tool wears and measuring the corresponding tool changes, the tool condition could be incorporated into the model.

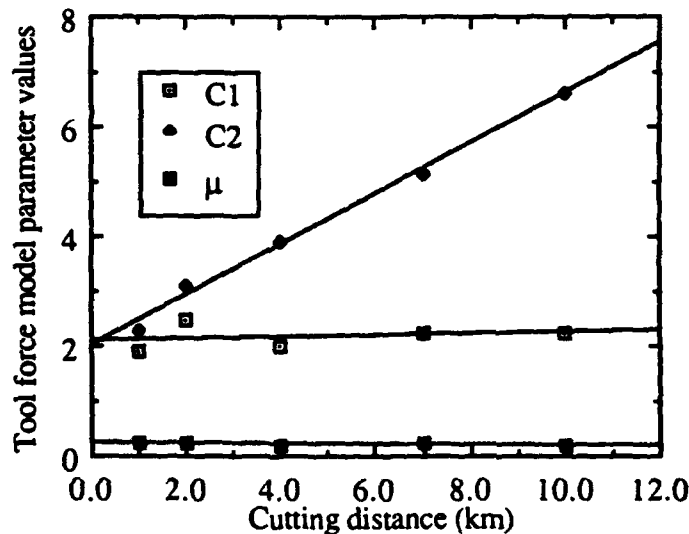


Figure 4: Parameters of the model (Equations.1&2) vs cutting distance.

The model described by Equations (1) & (2) includes forces at the tool edge which are not associated with material removal. These forces can be related to the extent of wear at the tool flank face. Figure 5 is a schematic of a tool on which a flat area has been worn. By assuming that the stress is the same on the rake face as it is near the clearance face, the wear width, w , can be predicted based on measured forces. Since the material in the vicinity of the tool edge has undergone extensive plastic deformation, a constant stress on the plastic zone boundary (the tool edge) is assumed. With measured force data and the fitting procedure, the parameters C_1 and C_2 are found. C_1 is the stress acting on area A_1 and the normal force on the area A_2 can be written as $C_1 L w$ by the constant stress assumption. Therefore:

$$w = C_2 / C_1 \quad (3)$$

Equation 3 states that the extent of flank wear can be calculated based on parameters derived from experimental force data. Equivalently, the flank wear can be included in the model for tool forces. The force model of Equations (1) and (2) can be written as:

$$F_Y = C_1 (L d + \mu L w) \quad (4)$$

$$F_Z = C_1 (\mu L d + L w) \quad (5)$$

The parameter C_2 in the model has been replaced by the new parameter, w , which has physical significance. Force data can be used to find the wear width, w , which best fits the data exactly as was done for C_2 , and the wear width can be checked by careful measurement of the tool.

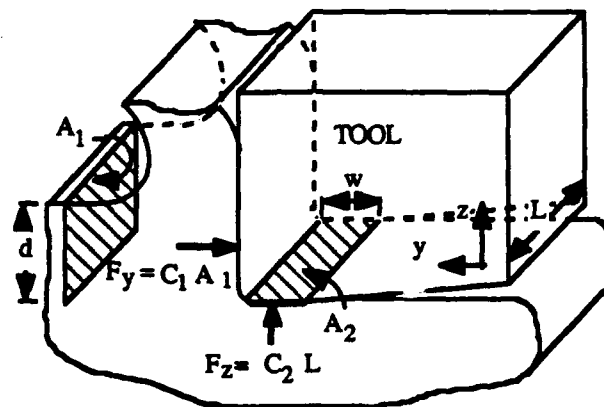


Figure 5: Tool edge diagram with edge wear

Measured forces were used to predict the wear width in two separate series of wear tests. In the first series, a straight-edge tool was used to cut plated copper for a total of 28 km. Periodically, force measurements were made over a range of cutting depths and the best fit parameters were found using the model of Equations (4) and (5). The values of w are plotted as '+'s in Figure 6. The width of the wear land as predicted by this procedure suggests that tool wear during the force measurements was small.

A second series of tests was performed by cutting Al 6061-T6 to accelerate the tool wear. A comparison of wear rates between aluminum and copper was also achieved in the process. The same tool was used after relapping. Force measurements were made using the same procedure as the first series of tests. That is, while aluminum was used to accelerate wear, forces were measured cutting the same hard, plated copper at specific stages of tool wear. The calculated values for w were obtained and are plotted in Figure 6 as the open circles. Note that after 1 km of cutting the aluminum, wear is greater than that after 28 km of cutting the hard copper.

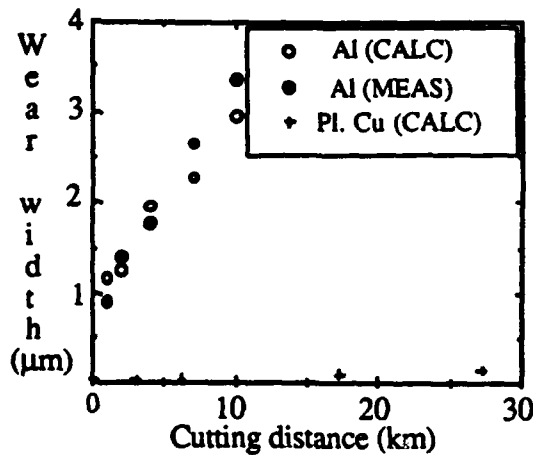


Figure 6: Calculated and measured tool wear vs. cutting distance

At each stage of wear in the second series, the tool was examined in a field emission scanning electron microscope (FESEM). The tool was imaged uncoated and 20 000X images were obtained at 5 keV using digital acquisition at imaging scan rates to avoid charging. A sample of the SEM photos used for the edge measurements is given in Figure 7. Figure 7(a) provides a schematic of the tool orientation in the SEM. It was necessary to know this fixturing geometry to obtain length measurements from the photos. Figure 7(b) is the micrograph taken of the tool after 2 km of cutting aluminum. Lapping marks are visible on both the clearance face and rake face of the tool. These have been worn away at the edge as the flat area developed. The wear dimension of interest is perpendicular to the rake face. Wear lines can be seen along the cutting direction on this wear area.

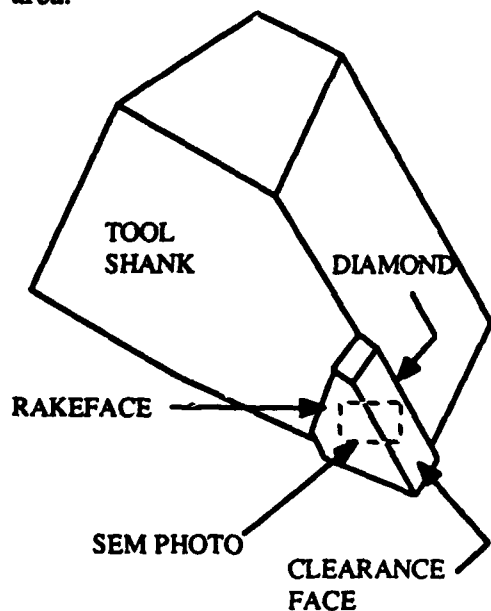


Figure 7(a): Top view of straight edge, diamond tool orientated in SEM

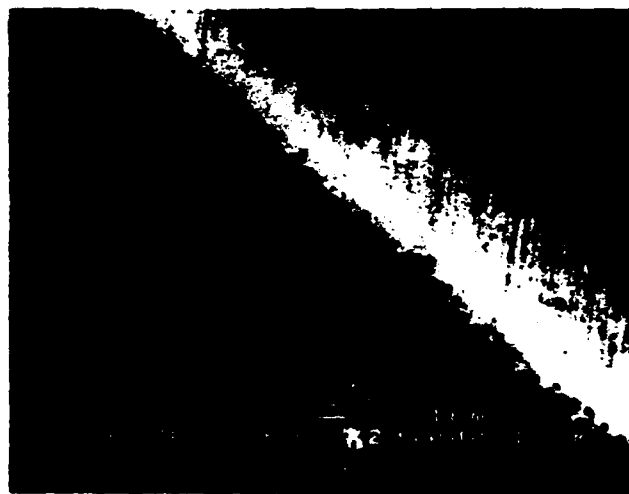


Figure 7(b): Micrograph of tool edge after 2km of cutting Al 6061-T6

The measured values of the wear width are plotted in Figure 6 as the closed circles. These values are in agreement with the wear predicted from the tool forces. The tool edge condition was not measured for the initial test series. However, the agreement of the measured and calculated tool wear for the second series indicates that using the force model to calculate wear width is a reliable procedure. The wear rate cutting aluminum was approximately 200 times greater than the wear rate cutting the plated copper.

Each calculated point in Figure 6 represents force data vs. depth of cut similar to Figure 3. Therefore, forces from both series of tests can be put on a single plot to show force vs. wear. The data covers an increase in wear land width of 100 times from 41 nm to 3 μm . Figure 8 is a plot of the normalized force components (Force / edge length) as functions of depth of cut and wear width. The cutting force is most influenced by depth of cut and to a lesser degree by tool wear while the thrust force is dominated by the tool edge condition. The ratio of measured thrust force to measured cutting force changes continuously along a line of constant cutting depth. Therefore, using the model represented by Figure 8, a single force measurement has enough information to assess the tool condition.

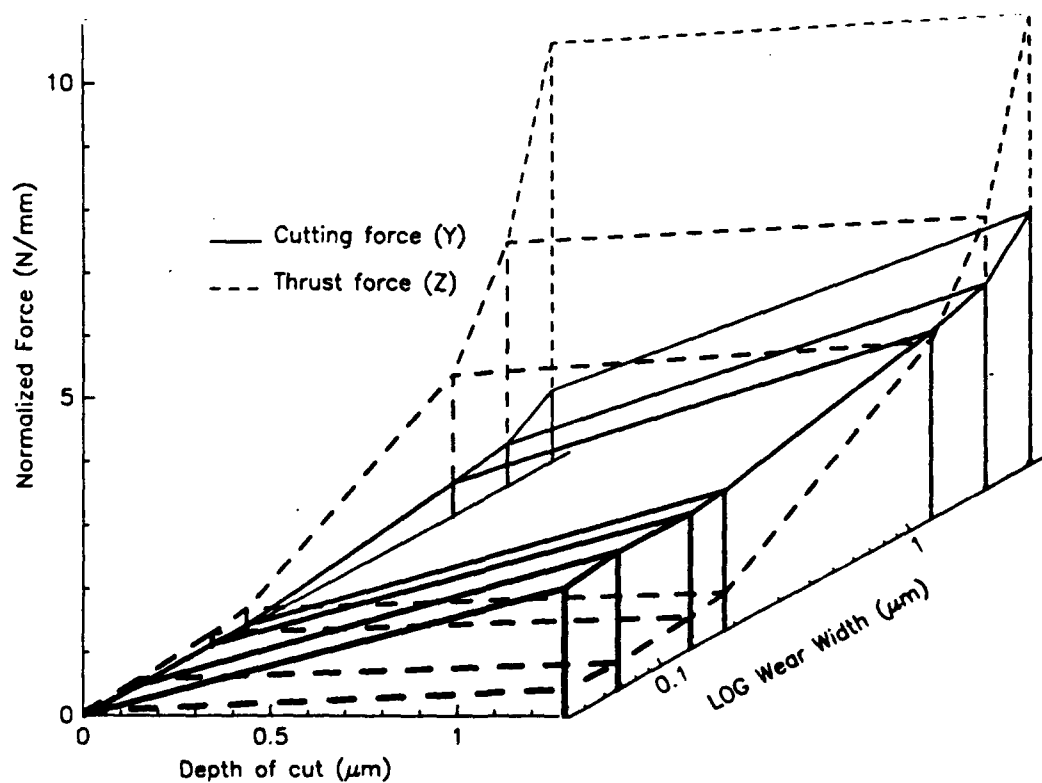


Figure 8: Tool force vs. edge wear and depth of cut using the model of Equations (1) and (2) with parameters derived from experimental data at the various wear stages

7.2.5 Material effects

The remaining aspect of the tool force model was to relate the parameter, C_1 , to some workpiece material property or properties. This was important to achieve the goal of developing general relationships. Although C_1 was found to be constant with respect to tool wear, Carroll's work suggests that this value would be material dependant [1]. In testing of other workpiece materials, this was indeed found to be the case. Significant effort was spent analyzing the relationship between C_1 and material.

Flow stress The terms flow stress and dynamic shear stress have been used interchangeably in metal cutting analyses. In metal cutting, flow stress is defined as the cutting forces resolved onto the shear plane divided by the shear plane area. Traditionally, the frictional forces on the clearance face of the tool are considered negligible so that in terms of measured cutting force, F_c , and thrust force, F_t , the flow stress can be written:

$$\tau_s = \frac{F_s}{A_s} = \frac{F_c \cos\phi - F_t \sin\phi}{A_0 / \sin\phi} \quad (6)$$

where:

- ϕ = shear plane angle
- F_s = shearing force on shear plane
- A_s = shear plane area
- A_0 = cross-sectional area of uncut chip

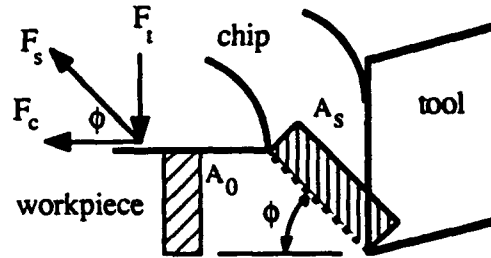


Figure 9: Forces in the shear plane

The dynamic shear stress, or flow stress, has been shown to be a material dependant constant for wide ranges of cutting parameters. Using the nomenclature of Figure 2, the flow stress can be written in terms of only forces acting on the rake face.

$$\tau_s = \frac{F_s}{A_s} = \frac{F_{Yr} \cos\phi - F_{Zr} \sin\phi}{A_0 / \sin\phi} \quad (7)$$

which can be simplified using the relationships $A_0 = Ld$ and $F_{Zr} = \mu F_{Yr}$.

$$\tau_s = \frac{F_{Yr}}{Ld} [\sin\phi \cos\phi - \mu \sin^2\phi] \quad (8)$$

In terms of the model parameter, C_1 , where $C_1 L d = F_{Yr}$:

$$C_1 = \frac{\tau_s}{\sin\phi \cos\phi - \mu \sin^2\phi} \quad (9)$$

C_1 can be expected to stay constant if the geometrical factor, $1 / (\sin f \cos f - m \sin^2 f)$ does not change. This depends on the shear angle alone if the friction coefficient remains constant. The theory of Merchant [4] predicts a shear angle of 40° using $f = 45 - b / 2 + g / 2$

b = friction angle = $\tan^{-1}(m)$ (11° with $\mu = 0.2$)

g = tool rake angle (0° used here)

Chips from the cutting tests were collected and measured in the SEM to calculate f from the ratio of cutting depth to chip thickness. If these are equal, the angle is 45 degrees. The uncut chip thicknesses (cutting depths) were 0.3, 0.6, and 1.2 μm . Examinations at 20 000X yielded chip thicknesses from which shear angles of 39° , 41° , and 39° were calculated respectively. A variation of shear angle from 30° to 50° in Equation (9) only produces a variation from 2.4 to 2.7 in the geometrical factor relating C_1 to t_s . Therefore, it is not surprising that C_1 was found to be constant for depth of cut and tool wear. C_1 can be considered as a material property related to flow stress by a constant.

The value of C_1 can be estimated for any material with force data from only two cuts if the depth of cut is varied. Equation (4) can be written for each cut and these can be combined into one equation, $(F_Y)_2 - (F_Y)_1 = C_1 (L d_2 - L d_1)$ where only C_1 is unknown. The uncertainty in this calculation can be reduced by using data from several cuts covering a range of cutting depth.

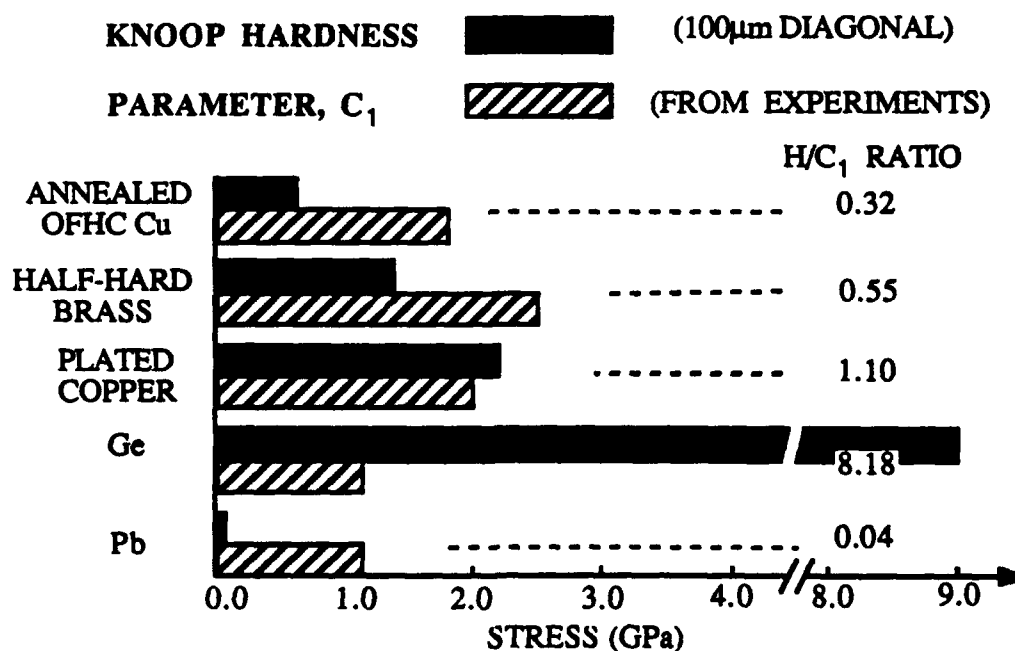


Figure 10: Knoop hardness and flow stress for various materials

Indentation hardness Indentation hardness was the first material property used for comparison. To find if a relationship between flow stress, C_1 , and hardness exists, several different workpieces were prepared and cut. Measuring forces enabled the flow stress to be calculated for each material as just described. The parameter C_1 was obtained for the electroplated copper, OFHC copper, and 70/30 brass. Both C_1 and Knoop hardness for each material are shown in Figure 10. Two other materials are included, single crystal germanium and lead, Pb. The ratio of indentation hardness to C_1 was calculated for all five materials. This ratio is included in Figure 9 in the column to the right of the bargraph. For these coppers and copper alloy the values acquired were not related to measured Knoop hardness.

Specific Cutting Energy Force in the cutting direction divided by uncut chip cross-sectional area, is commonly called specific cutting energy [5]. The cutting force multiplied by cutting velocity is power and cross-sectional area multiplied by velocity is volumetric removal rate. The ratio of these is force per unit area and also work per unit volume. Therefore, cutting force can be a measure of energy spent per unit of material removed in the process.

It is common to use the measured force in the cutting direction in the calculation of specific cutting energy. But this value includes the frictional force on the tool edge contact area. Depth of cut and tool edge condition are factors. At small depth of cut, a greater percentage of work is associated with the friction force at the tool edge which does not involve material removal. Also a dull tool causes more frictional work than a sharp tool for the same cutting depth. It is not uncommon, therefore, for a range of specific cutting energy values to be reported for a single material. However, C_1 can be interpreted as a specific energy calculated without contributions to force from friction.

The interpretation of C_1 as volumetric work suggests it is related to the area under the stress strain curve. It is the work necessary to deform the material both elastically and plastically to the point of failure. This area is a function of both the strength of the material and its strain at separation. The latter is really ductility. Figure 11 is a schematic showing this interpretation of the flow stress, strength, ductility relationship. Since metals do not generally have both high strength and high ductility, it is possible for the area under the stress-strain curve to be larger for a softer material. Since indentation hardness is mainly a function of strength, the poor relationship between hardness and flow stress, or C_1 , can be understood. The data for Ge and Pb were included in Figure 10 as extremes to emphasize the importance of both ductility and strength to the measured flow stress. Germanium is very hard but brittle. In contrast, lead is soft and very ductile. It is the combination therefore which results in nearly equal values of C_1 .

The second attempt at incorporating material properties into the force model involved the stress-strain characteristics. A shearing experiment was chosen to test three materials, (plated copper, OFHC copper, and 70/30 brass), and obtain shear stress-strain curves. This was considered to approximate the conditions in actual cutting better than a standard tensile test.

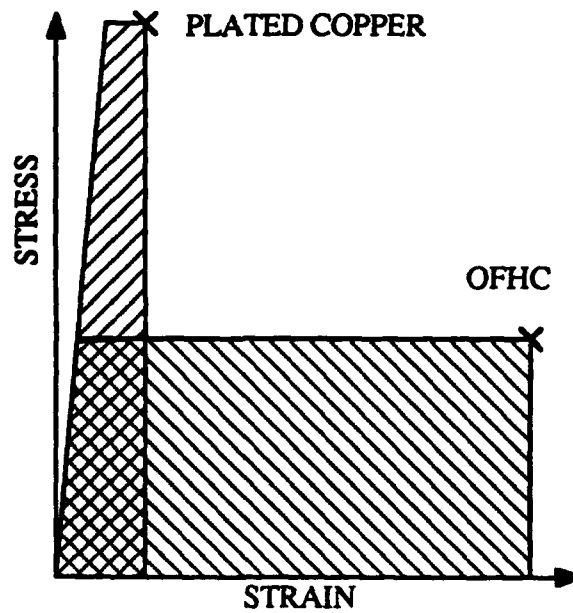


Figure 11: Schematic stress-strain curves

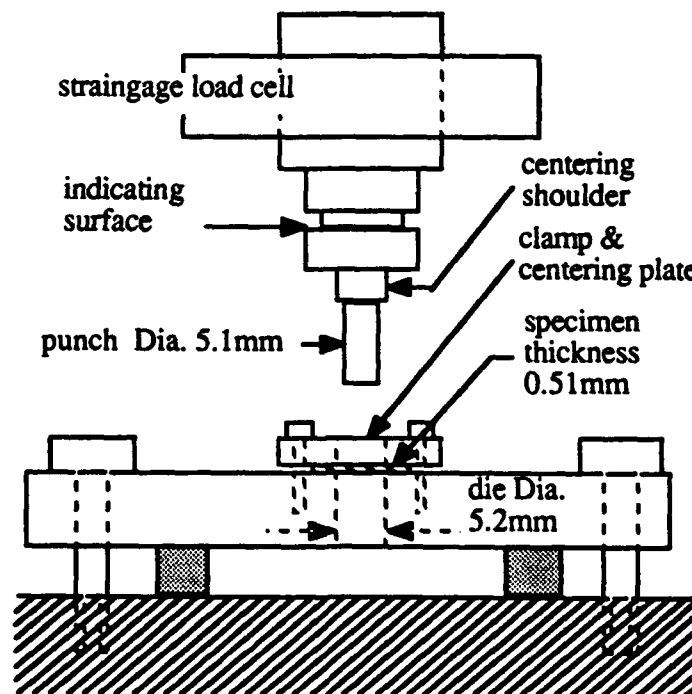


Figure 12: Schematic view of the punch and die shear testing apparatus used on thin plates of brass and copper

The apparatus for the material testing is shown in Figure 12. This was used on the 1000 lb tensile test machine in the PEC. It consisted of a punch and die made of hardened steel. The die was bolted to the base of the testing machine. The punch was mounted to the strain gage load cell on the moving cross member. A slotted ring on the punch allowed for a Federal gage indicating needle to measure displacement as the material was punched. Force was recorded simultaneously from the output of the strain gage. The test specimens were 500 μm thick and the die-punch radial clearance was 65 μm or 13 % of the thickness. There was a shoulder on the punch with a tight fit in the die. This served to center the punch relative to the die before the die was tightened to the machine base.

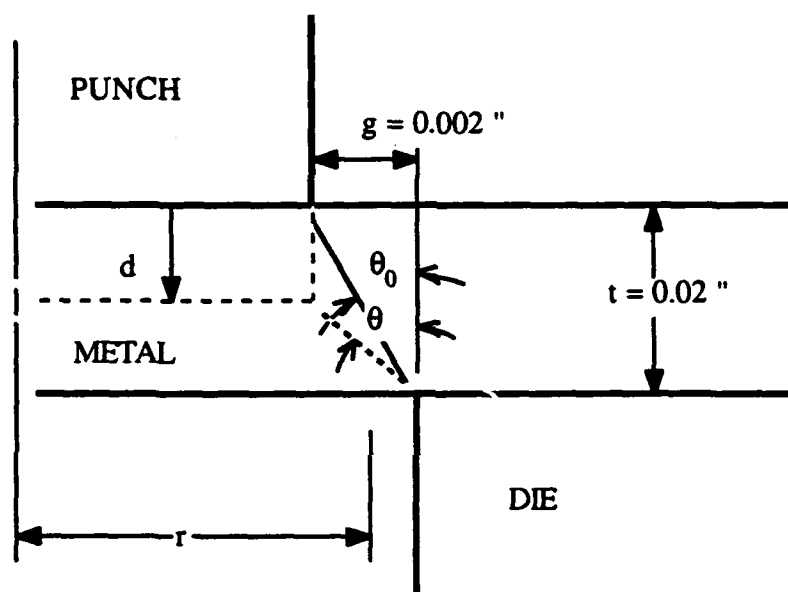


Figure 13: 2-D model for shear stress-strain calculations.

Shear stress and shear strain were calculated from the force and displacement data using pure shear assumptions as the 2-dimensional model of Figure 13 shows. The shear stress, τ , is the force divided by shear area. This area is the circumference multiplied by the current load bearing thickness which is the original sample thickness minus the punch penetration. The shear strain is the integral of the change in angle at the same punch penetration depth. The shear stress and shear strain equations are given as:

$$\tau = \frac{F}{2\pi r(t-d)} \quad (10)$$

The three curves generated using equations (10) and (11) are superimposed in Figure 14. The plated copper had the highest yield point as expected. This is the stress value where nonlinearity first appears. The brass exhibited extensive work hardening characterized by flattening of the curve with increased stress at strains above yielding. The annealed OFHC copper had the highest ductility as seen by the strain at failure.

The areas under the curves of Figure 14 were obtained by numerical integration of the data. These values were compared to values for C_1 from the cutting experiments. Figure 15 is a bargraph similar to the one comparing hardness to flow stress. Although flow stress and stress-strain area do not match exactly, the ratios are similar for all three materials. The difference in temperature and strain rate between cutting and the shear test may be the cause of the deviations. Since all three materials are similar, they are all affected in the same way. That is, the values from shear test are all higher than measured flow stress.

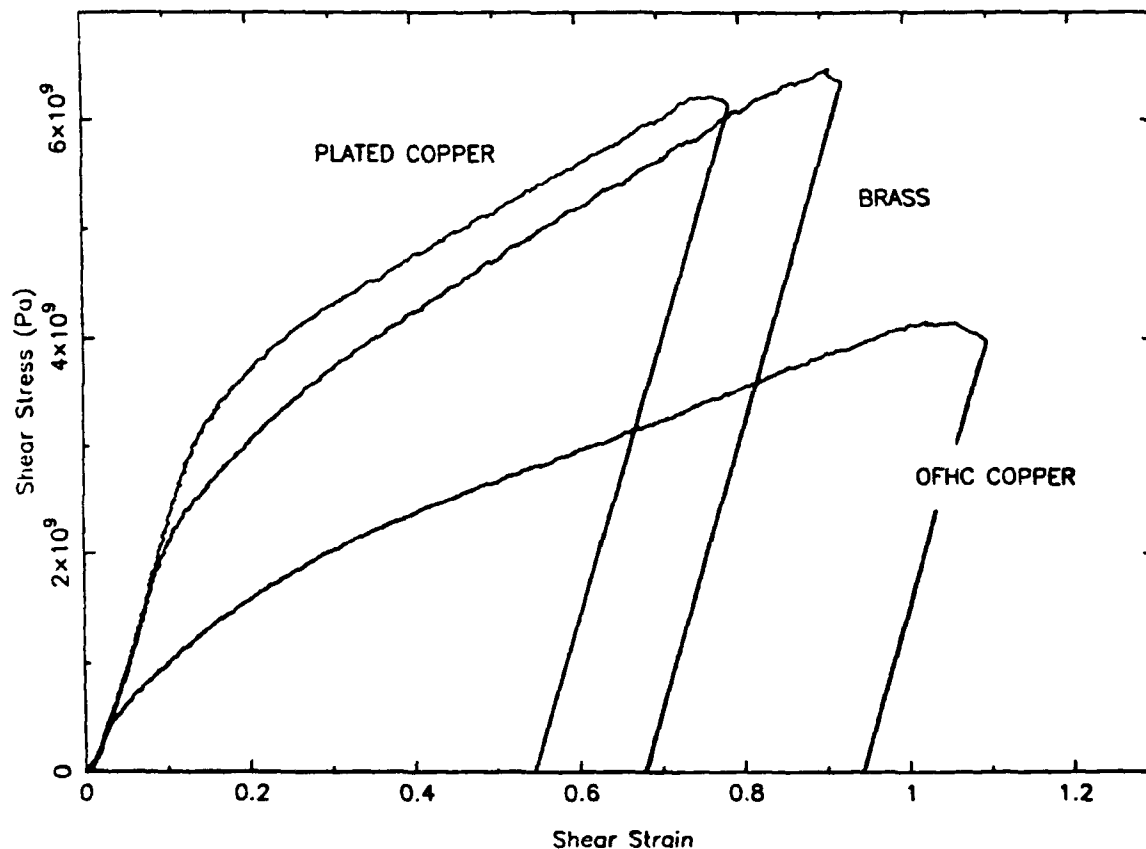


Figure 14: Shear stress-strain curves for plated copper, half hard brass, and annealed OFHC copper, constructed from force-displacement data.

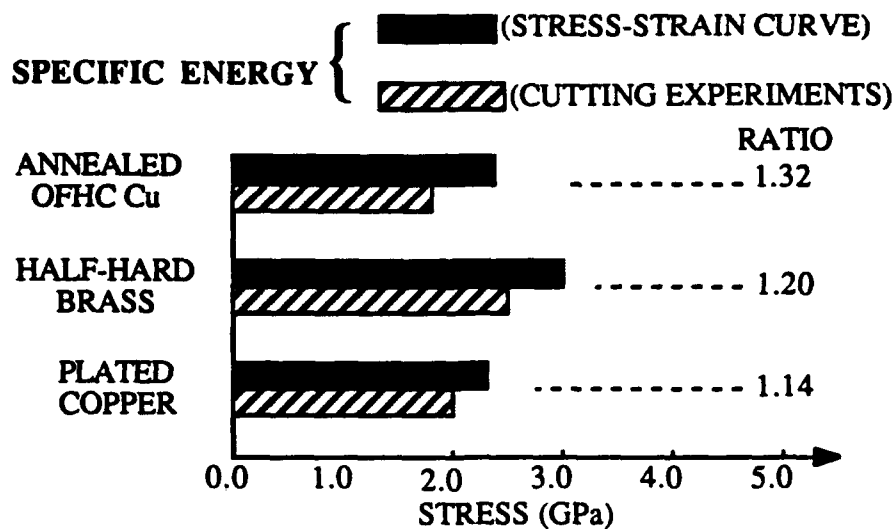


Figure 15: Flow stress from force measurements compared to energy calculated from shear stress-strain curve areas.

7.3 APPLICATION OF MODEL TO ROUND-NOSE TOOL GEOMETRY

An important step in this research is to apply the force model derived from ideal, orthogonal cutting tests to the round-nose finishing geometry. The goal is to relate tool force to the surface produced with the finishing tool. Therefore, it must be verified that forces can be predicted in this slightly different process.

7.3.1 Force prediction for finishing operation

It is fairly easy to adapt the equations for tool force to the geometry of the cutting tool used in finishing. Figure 16 shows the geometry of the cutting cross-section. In the round-nose tool cutting geometry the actual cutting depth, or chip thickness, varies from zero near the bottom of the cut to a maximum over the contact angle of the tool. To apply the results of Figure 1, the chip cross-section can be divided into a number of small segments, where the thickness of each can be considered a constant. The normal force on the rake face is then the flow stress, C_1 , multiplied by the summation of the cross-sectional area. The frictional force on the rake face and the normal force on the clearance face wear land are assumed to act normal to the rounded cutting edge.

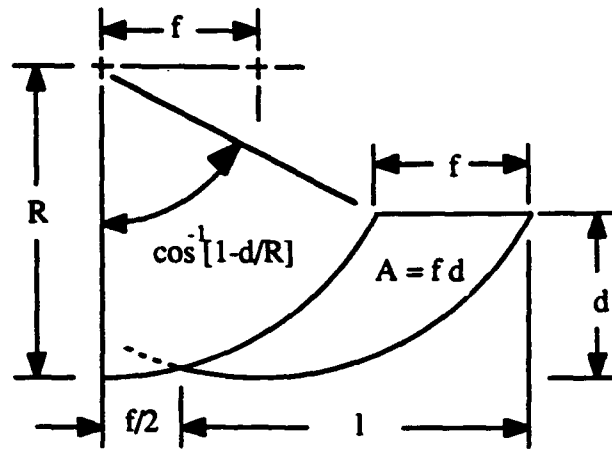


Figure 16: Cross-sectional view of the round-nose tool cutting geometry.

Looking at the geometry of Figure 16 and using the nomenclature of Figure 2,

$$F_{Yr} = C_1 f d \quad (12)$$

$$F_{Zr} = \mu C_1 f d \quad (13)$$

$$F_{Ze} = C_1 w l \quad (14)$$

$$F_{Ye} = \mu C_1 w l \quad (15)$$

In these equations,

d = overall depth of cut

f = infeed / revolution

w = edge wear land width

l = length of contact projected into the Z direction = $R \sin[\cos^{-1}(1-d/R)] + f/2$ where

R = the tool nose radius

The equations for measured forces are then the sums of the components of Equations (12-15)

$$F_Y = C_1 [f d + \mu w \{ R \sin[\cos^{-1}(1-d/R)] + f/2 \}] \quad (16)$$

$$F_Z = C_1 [\mu f d + w \{ R \sin[\cos^{-1}(1-d/R)] + f/2 \}] \quad (17)$$

To apply this model, the force parameters were first determined from force measurements using straight-edge tools. Figure 17 shows a typical curve from which the parameter values were obtained. In this case the material was OFHC copper and the values for the flow stress, friction coefficient, and wear land width are 1.82GPa, 0.32, and 1.29 μ m respectively.

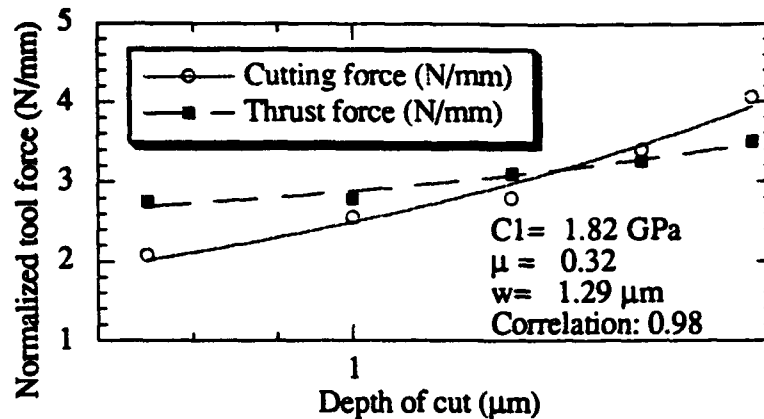


Figure 17: Experimental cutting data from straight-edge geometry from which parameters are derived. These are used in Equations (16) and (17) to predict forces in the round-nose geometry of Figure 16.

The force curves were calculated using Equations (16) and (17) and then plotted as shown in Figure 18. In this figure, depth of cut, tool nose radius, and wear width were input. The forces are plotted vs. feedrate. These calculated curves could then be compared to measurements of force in the round-nose cutting geometry.

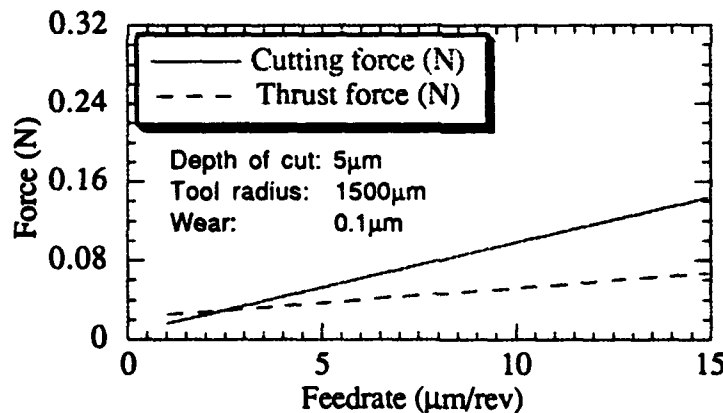


Figure 18: Predictions of cutting and thrust force vs. infeed for the round-nose geometry. These curves are generated using parameters derived from the straight-edge cutting tests.

7.3.2 Correlation with cutting tests

A series of cutting tests was performed in which the tool was used to cut OFHC copper for a total distance of 460km. At 14 different stages of wear, force measurements were made for a range of infeeds and the tool was examined. After 128km of cutting, the tool exhibited a wear width of 0.085μm. The forces measured at that wear stage were superimposed on the same curves as

Figure 18. The measured forces are shown in Figure 19 as data points. Although the calculated and experimental values were in agreement, the application of strictly orthogonal cutting geometry to the geometry of a round-nose tool is not as straightforward as originally expected and this topic needs further investigation.

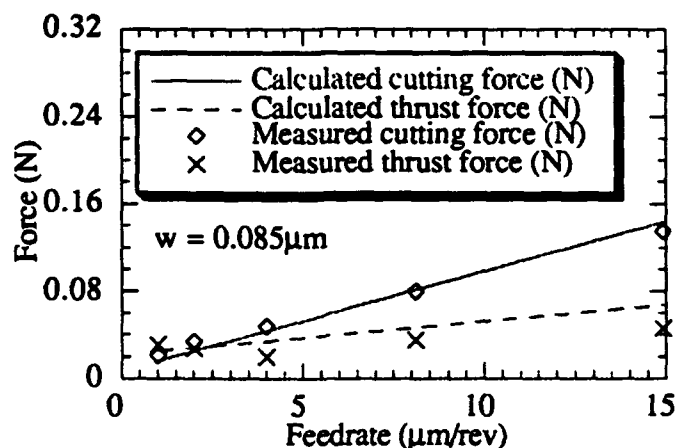


Figure 19: Reproduction of Figure 18 with the addition of data points from round-nose cutting experiments. These forces, shown as data points, were obtained after 128km of cutting copper. The wear width at this point was approximately equal to the $0.1\mu\text{m}$ used in the calculated curves.

There are obviously differences between the two geometries of straight edge and rounded nose tools and relating tool forces between the two may require a close examination of these differences. One difference seems to be related to the depth of cut. With the straight-edge tool where the infeed is equal to the depth of cut, this depth is known by setting the spindle speed and the machine slide velocity. In the finishing geometry, however, the depth of cut can only be estimated after the part is removed from the spindle and measured. If there is burnishing taking place in the process, or elastic springback, it is difficult to tell to what extent this has occurred. The point on the arc defined by the assumed edge radius where the chip separates from the workpiece may be an important parameter. This point is also important to the surface produced in the finishing process. Therefore, an understanding of the mechanisms responsible for surface roughness may help in the solution for predicting force in the finishing geometry.

7.4 SURFACE FINISH MODELING

A model for surface finish which includes tool and material parameters is equal in importance to a force model if the surface is to be related to force. It is well known that surface finish in diamond turning (DT) is related to the condition of the tool. Skill and experience has been applied to the finishing of tools for the best possible surface finish. The difficulty involved in making tools

capable of optical finishing is evident in their cost. The price can be as much as 100 times the cost of steel or carbide tooling used in conventional material removal. It is important to understand wear of diamond tools and particularly how the condition of the tool edge affects the surface finish which is the primary concern in DT applications. The previous section showed the feasibility in quantitatively assessing the tool edge condition using force information from the process. Because the tool is related to the finish, the possibility is raised for assessing the surface finish quality as it is being produced.

7.4.1 The role of vibrations in surface roughness

A preliminary experiment was conducted to determine the effect of tool wear on surface finish. In conjunction with the series of cutting tests in which forces were measured at various stages of wear, sample surfaces were also produced. Surface characteristics such as R_a , RMS, and R_{p-p} were measured for each surface. The surface finish was little affected by tool wear, while the changing forces showed that wear was occurring, it was concluded that the tool was not the dominant factor in determining roughness values. The machine vibrations, including spindle, slideway, and possibly tool post motion were dominating the results. An extreme example is presented in Figure 20. This is a Zygo Maxim 3-D measurement of a DT surface. The three surface parameters are shown at the left. Peak to valley roughness is given as 69nm and a spatial periodicity of between 15 and 20 μ m is evident. The infeed used for this cut was 5 μ m/rev. This means that a dominant machine vibration existed during the cutting which came in phase with the spindle revolution frequency approximately every 4.5 revolutions. There are many possible vibration frequencies which can be calculated which would cause this surface anomaly. However the amplitude of the vibration, 70nm, limits the possibilities to the lower frequencies and one possibility is 72 Hz.



Figure 20: Topographic data of a copper surface diamond turned with a 5 μ m/rev infeed. The apparent periodicity which is larger than 5 μ m shows a phase condition between machine vibration and the spindle revolution frequencies.

Figure 21 shows the frequency response of the motion between spindle and tool post in a direction parallel to the spindle axis. This data was acquired using a capacitance gage while the spindle was not turning. A large, broad peak can be seen at the cursor placement of 72.5Hz. Other spikes at 60Hz and its harmonics show electrical noise in the measurement. A measurement of motion amplitude using the same measurement set-up revealed a consistent 40-50nm value. The conclusion was that imperfections in surface finish are caused by vibration between tool and workpiece.

Although it is not the intent of this research to investigate machine vibrations, it was necessary to reduce their effects so that the contributions of tool edge to surface imperfections could be studied. A systematic approach of investigating the cause of the vibrations was carried out. This included various fans, motors, and lab noises in their on and off conditions.

The most effective reduction in vibration occurred when the spindle axis slide was allowed to rest on the machine base. This was accomplished by installing a shut-off valve in the hydraulic bearing oil supply line to the z-slide. The control to the z-axis was also disabled so that there was no chance of the slide moving without its bearing. Although this is an impractical solution in general, it allowed turning of flats where control of cutting depth was not critical.

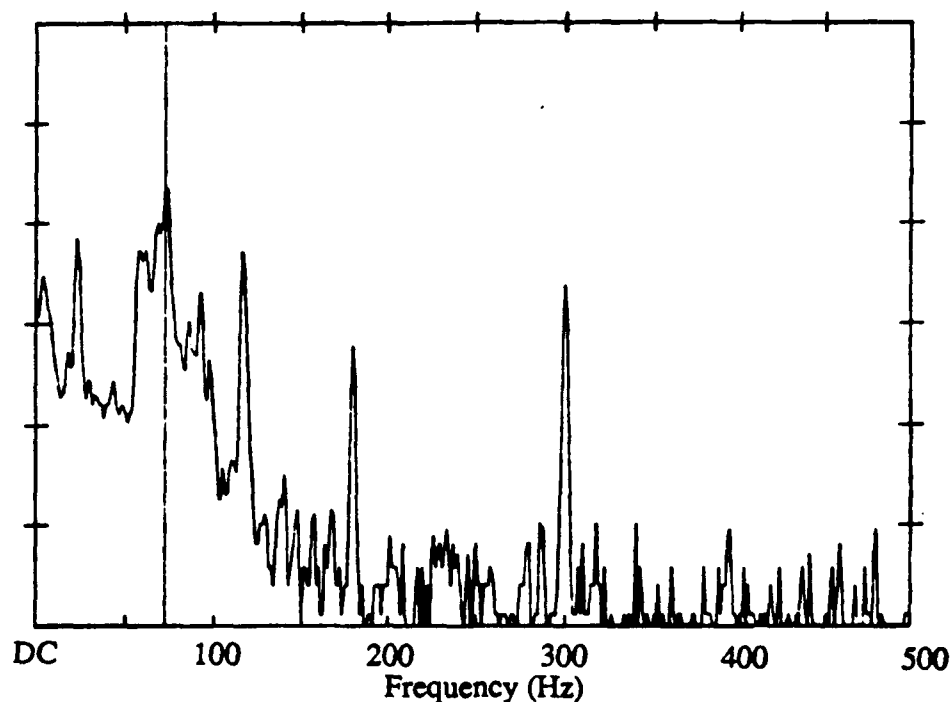


Figure 21: Frequency response of the motion between spindle and capacitance gage mounted on the tool post of the diamond turning machine measured by a Wavetek frequency analyzer.

The frequency response acquired with the z-slide disabled is shown in Figure 22. The 70Hz vibration was substantially reduced. With surfaces produced in this way surface effects within a single feed, or several adjacent feeds could be examined.

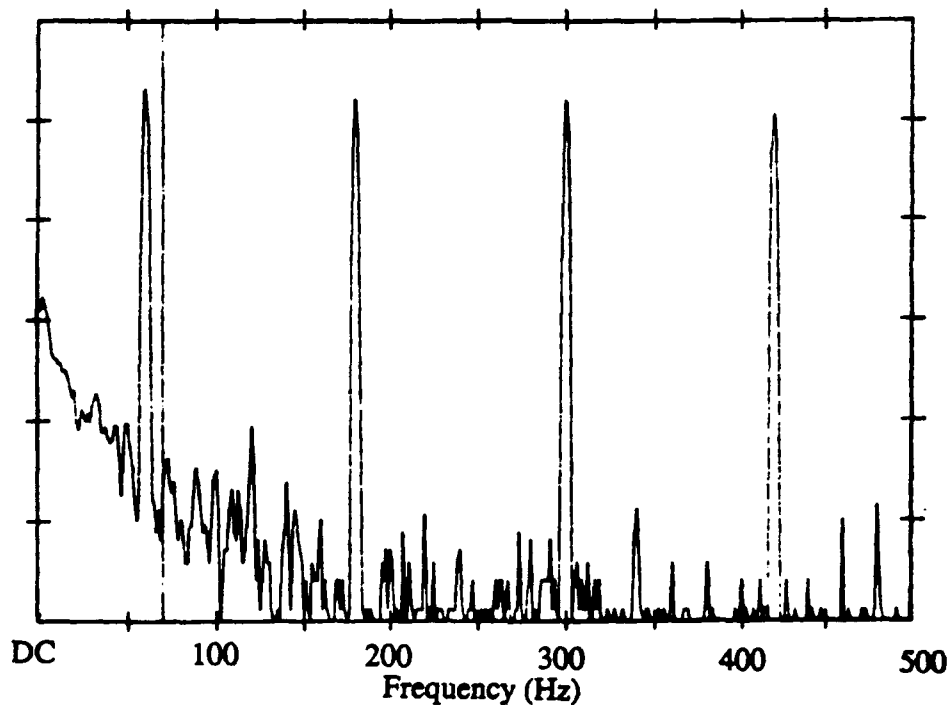


Figure 22: Frequency response of the spindle relative to the tool post after allowing the spindle slide to rest on the machine base by shutting off the hydraulic bearing fluid. Compared to the response in Figure 20, the vibration at 70Hz is greatly reduced.

7.4.2 Simulation of the theoretical surface

With the machine vibrations minimized it was possible to examine surfaces on a scale where they could be compared to the theoretical finish. Furthermore, it was found that the peak to valley excursions in a measurement could be reduced by choosing a line scan mode. By searching an area it was possible to find sections with several feed marks lying in a plane. Figure 23 shows one such measurement. The line scan is perpendicular to the tool marks which are in the cutting direction. It is clear that vibrations do not cause this consistent p-v roughness of 12nm. However, a theoretical roughness calculated for the 750 μ m tool nose radius and the 5 μ m/rev infeed is 4nm, that is, a third of the measured value. A detailed understanding of the tool material interactions in the cutting process is needed to explain this difference.

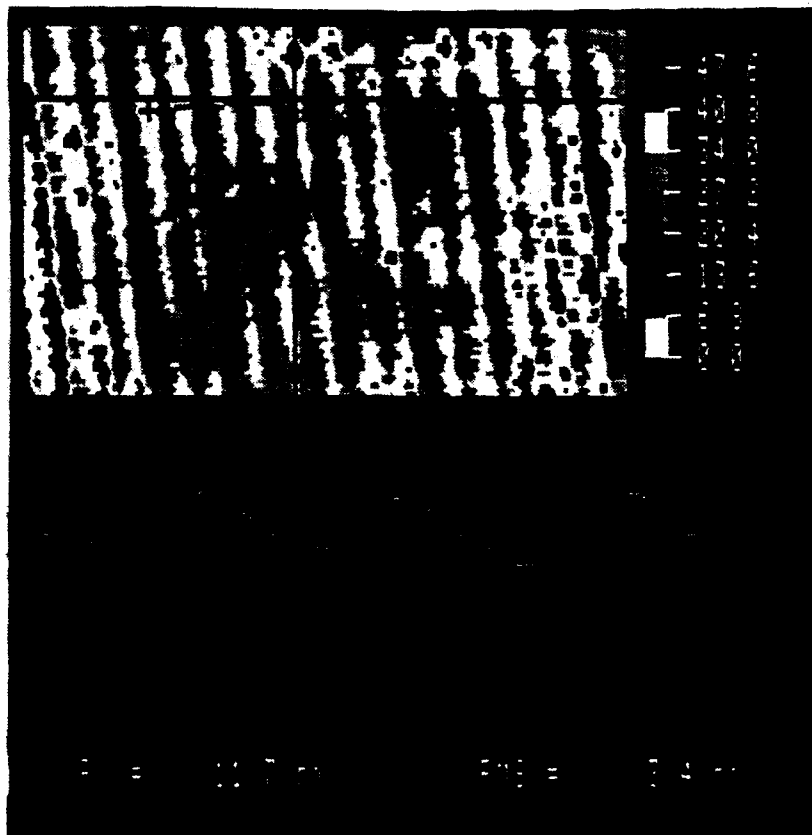


Figure 23: Zygo data from line scan measurements of a diamond turned copper surface in an area where the variation between successive tool passes was minimal.

To understand how a DT surface deviates from the theoretical finish, a plotting routine was written to simulate a line scan across the tooling grooves. The first step was to generate the theoretical surface. The algorithm involved graphically overlaying several x-y plots of the tool edge. The tool was assumed to be perfectly round with a nose radius of $750\mu\text{m}$. Each successive arc was shifted in x by the infeed spacing. The minimum y-value at each x coordinate was then selected to be plotted. Figure 24 is a plot of the ideal surface created using a $5\mu\text{m}$ infeed/revolution. A center line is also drawn and the parameters of interest are calculated. The peak to valley roughness is 3.85nm which is very close to the height calculated using the approximate formula,

$$R_{p-p} = \frac{f^2}{8r} \quad (18)$$

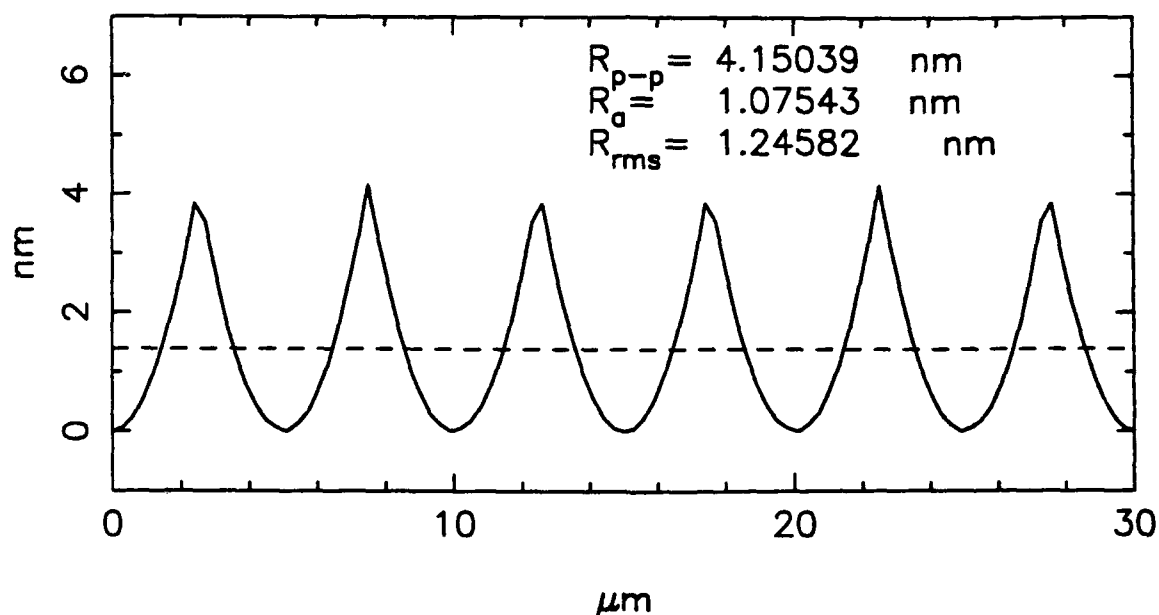


Figure 24: Simulation of an ideal diamond turned surface having an infeed of $5\mu\text{m}/\text{rev}$ and tool nose radius of $750\mu\text{m}$.

7.4.3 Addition of minimum depth of cut criteria

The actual cutting process is not ideal because the tool edge is not perfectly sharp. There is generally assumed to be a finite roundness to the cutting edge (See Section 9). The material cannot be removed cleanly as required in the theoretical model. The schematic of Figure 16 showed the cross-sectional area of the uncut chip. The chip becomes zero thickness at the trailing edge of the cut. Both the infeed and depth of cut are greatly exaggerated in Figure 16, distorting the scale, and in fact the thickness is very small along most of its length. At its thickest, the chip is about $0.6\mu\text{m}$. This is at a distance of around $130\mu\text{m}$ from the center of the tool nose. The chip thickness increases linearly from 0 to $0.6\mu\text{m}$ over a length of $130\mu\text{m}$.

To better approximate the actual cutting process, a minimum depth of cut criterion was added to the model. In the simulation, for chip thickness smaller than this minimum, no material was removed. At greater thicknesses, the material was assumed to be cut cleanly. This minimum depth of cut was expressed as a fraction of the tool edge radius. The result is shown in Figure 25.

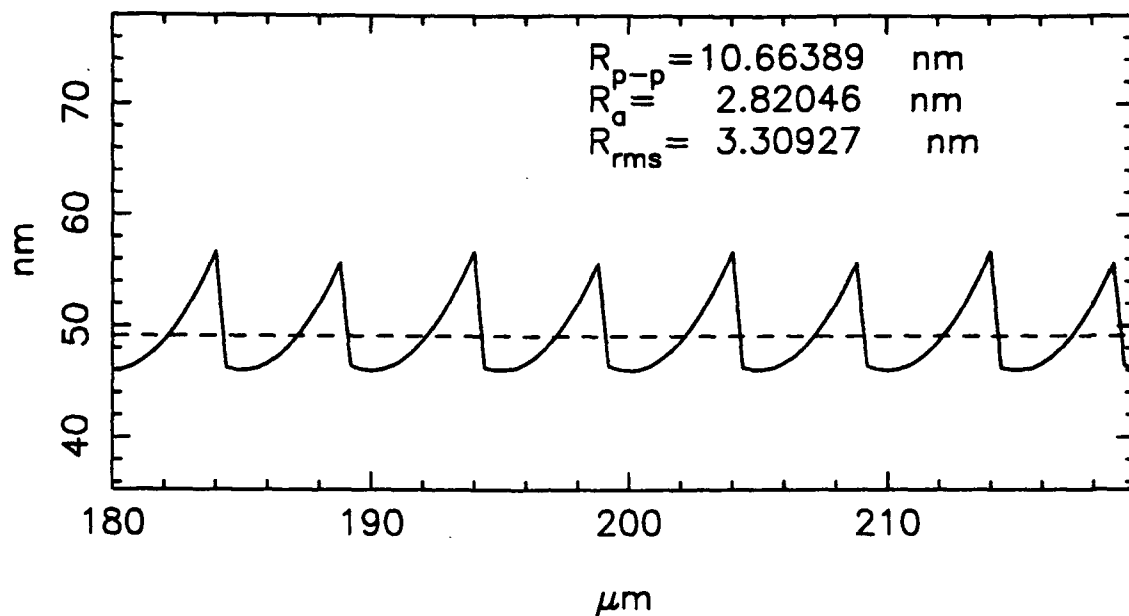


Figure 25: Simulation of a line scan across the tool feed grooves in which a minimum depth of cut criterion is established. The part of the chip having thickness less than this minimum is not removed. At greater thickness the chip is removed cleanly.

For this addition to the calculated surface model, an estimation of $0.2\mu\text{m}$ was used for the edge radius. Although a theoretical limit for sharpness has been reported to be 20nm [6], estimations from experimental evidence and direct measurements give values from 0.05 to $1.0\mu\text{m}$ [7,8].

The result of Figure 25 gives the correct magnitude of R_{p-p} roughness based on the measured surfaces but the shape is unrealistic when compared to the line scan of Figure 23. An abrupt change from burnishing to cutting is not realistic because the change in thickness along the chip is slow as previously discussed. The actual transition from a burnishing process to complete material removal could better be modeled as gradual.

A refinement of the surface model was to incorporate a second cutting depth parameter, d^* , greater than the minimum depth of cut, above which the material was cut cleanly. That is, below d_{\min} no chip is produced. Above d^* , the chip is separated with a thickness equal to the theoretical uncut chip thickness. d^* was also expressed as a ratio of the edge radius. Between d^* and d_{\min} a gradual transition in the form of a cosine curve was used. The resulting simulated surface is shown in Figure 26. The criteria used for this simulation were:

$$\begin{aligned} d_{\min} &= \text{minimum chip thickness for cutting} = 0.38 \rho \\ d^* &= \text{chip thickness for complete cutting} = 0.53 \rho \\ \rho &= \text{tool edge radius} = 0.2 \mu\text{m} \end{aligned}$$

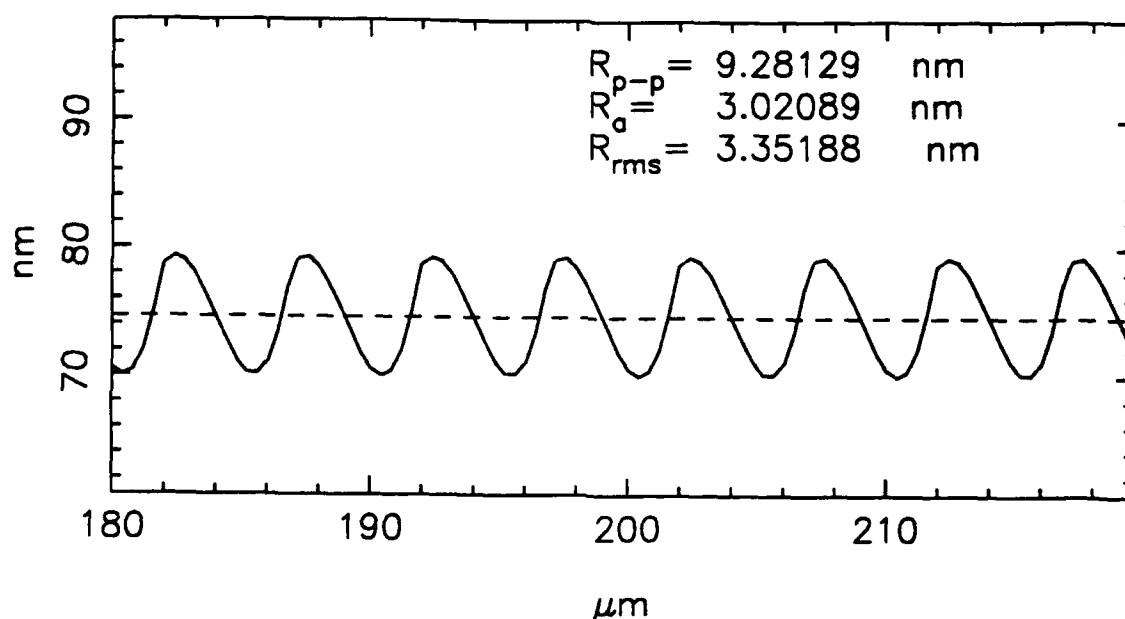


Figure 26: Simulation of a line scan across the tool feed grooves. In this simulation a minimum depth of cut criterion is used as well as a second cutting depth criterion. For chip thickness greater than this second cutting depth, material is removed completely. For thickness between these two values, some but not all material is removed.

Figures 27 and 28 show the effects of minor changes in the depth of cut criteria. If d_{\min} and d^* are increased in the simulation to 0.39ρ and 0.55ρ respectively, the result is Figure 27. Another similar increase to 0.40ρ and 0.56ρ respectively produces the result of Figure 28. This suggests that very minor changes in the cutting variables which affect d_{\min} and d^* can have a drastic influence on the shape of the surface on a scale of a few tool grooves.

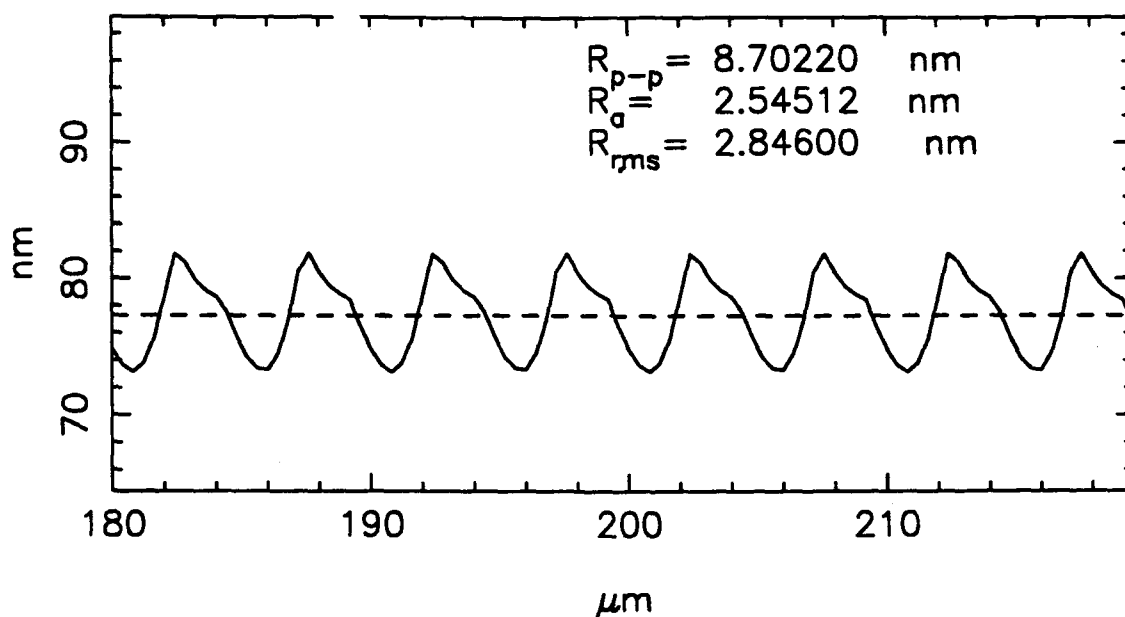


Figure 27: Simulation of a line scan similar to Figure 25 with the depth of cut criteria increased by 2.5%

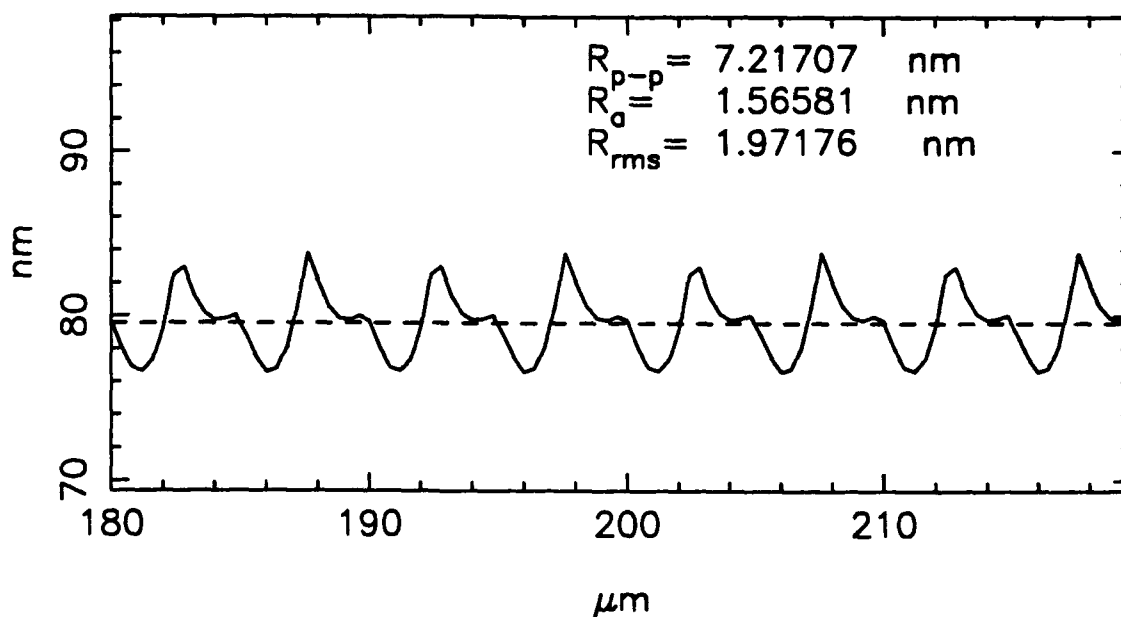


Figure 28: Simulation of a line scan similar to Figure 25 with the depth of cut criteria increased by 5%

7.4.4 Comparison to measured surfaces

The simulated surface plots show good agreement with the observed surfaces as measured with the Zygo optical interferometer, the Rank Taylor Hobson Talystep, and Digital Nanoscope STM. Figure 29 shows a particularly interesting Talystep line profile in which each of the three surface shapes of Figures (26-28) are evident. The apparent extra feed grooves of Figure 29 can be explained using the hypothesis of the minimum depth of cut. This similar shape is predicted by the simulation of Figures 27 and 28. If the minimum depth of cut occurs at a distance greater than the infeed from the tool centerline along the tool edge, the tool will encounter the same material more than once. The second time, the chip thickness adds to the material left behind and the d_{min} criterion is exceeded. Thus, the resulting surface anomaly occurs. To understand how this can happen, the scale of these figures must be kept in mind. The vertical magnification in the Talystep profile for example is 500 times greater than the horizontal magnification.

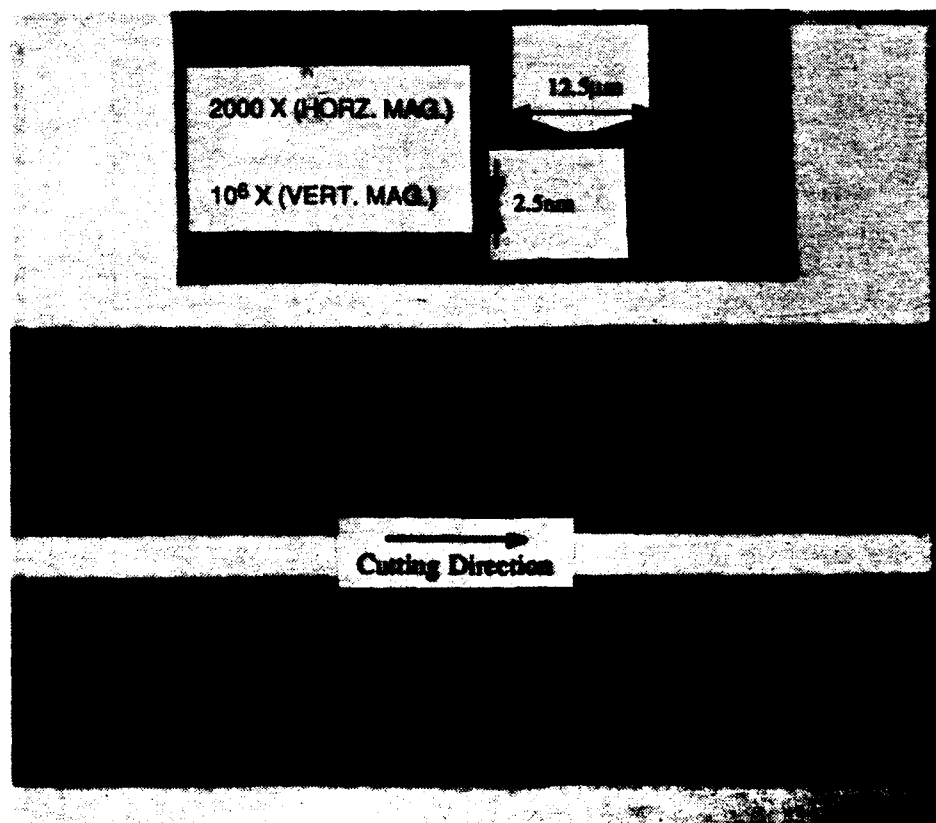


Figure 29: Talystep profile of a diamond turned surface which exhibits the same characteristics as the simulations of Figures 25, 26, and 27.

7.5 CONCLUSIONS

A model for tool forces in diamond turning of ductile metal was developed. This model quantitatively includes the wear state of the tool and properties of the workpiece material. The material properties important to tool force were found to be strength and ductility.

Application of the tool force model to the prediction of forces in typical finishing operations is an important step. There were unforeseen difficulties in making the transition from the orthogonal cutting model to one for a round-nose tool. The difficulties seem to be related to the chip cross-sectional thickness which is varying and very small in the round-nose tool geometry. More analysis and experimentation is necessary in this area.

Modeling has been initiated to describe a diamond turned surface. This model deals with the contribution of tool edge condition to the shape of the surface on the scale of the feed lines. A minimum depth of cut has been proposed based on the tool edge radius. Because material is not removed when the uncut chip thickness is below this minimum, the model accounts for the differences between the actual and theoretical surface.

7.6 FUTURE WORK

The emphasis of this research for the coming year will be to:

- Further refine the force model to better describe forces in the round-nose finishing geometry.
- Develop the surface model by establishing minimum depth of cut criteria based on principles of mechanics and materials science.
- Show how measured surface characteristics change as a function of tool wear.

References

- [1] J.T. Carroll, T.A. Dow, J.S. Strenkowski, "Tool Force Measurement and Prediction in Diamond Turning", SPIE Vol.676, Ultraprecision. Machining. & Automated. Fabrication of Optics, 1986.
- [2] J.D. Drescher, T.A. Dow; "Tool Force Model for Diamond Turning", NCSU Precision Engineering Center Annual Report, Vol.7; 1989.
- [3] J.D. Drescher, T.A. Dow; "Surface Finish, Tool Force Relationships in Diamond Turned Copper", NCSU Precision Engineering Center Annual Report, Vol.8; 1990.
- [4] M.E. Merchant; "Basic Mechanics of the Metal-Cutting Process", Journal of Applied Mechanics, Vol.11,ppA168-A175Sept, 1944.
- [5] G. Boothroyd, W.A. Knight; Fundamentals of Machines and Machine Tools, 2nd Ed. McGraw-Hill,Inc , 1975.
- [6] N. Ikawa, S. Shimada; "Cutting Tool for Ultraprecision Machining", Proc. of the 3rd International Conference on Production Engineering; pp357-364; July,1977.
- [7] C. Evans, R. Polvani, M. Postek, R. Rhorer; "Some Observations on Tool Sharpness and Sub-surface Damage in Single Point Diamond Turning", SPIE,Vol.802,pp52-66; 1987.
- [8] J.D. Drescher, T.A. Dow; "Machining Forces in Diamond Turning of Plated Copper and Unplated Substrates",Proceedings of the ASPE Spring Topical Meeting: Metal Plating and Plating Fabrication; Tucson, AZ 1991

8 DIAMOND TOOL WEAR

William C. Larson

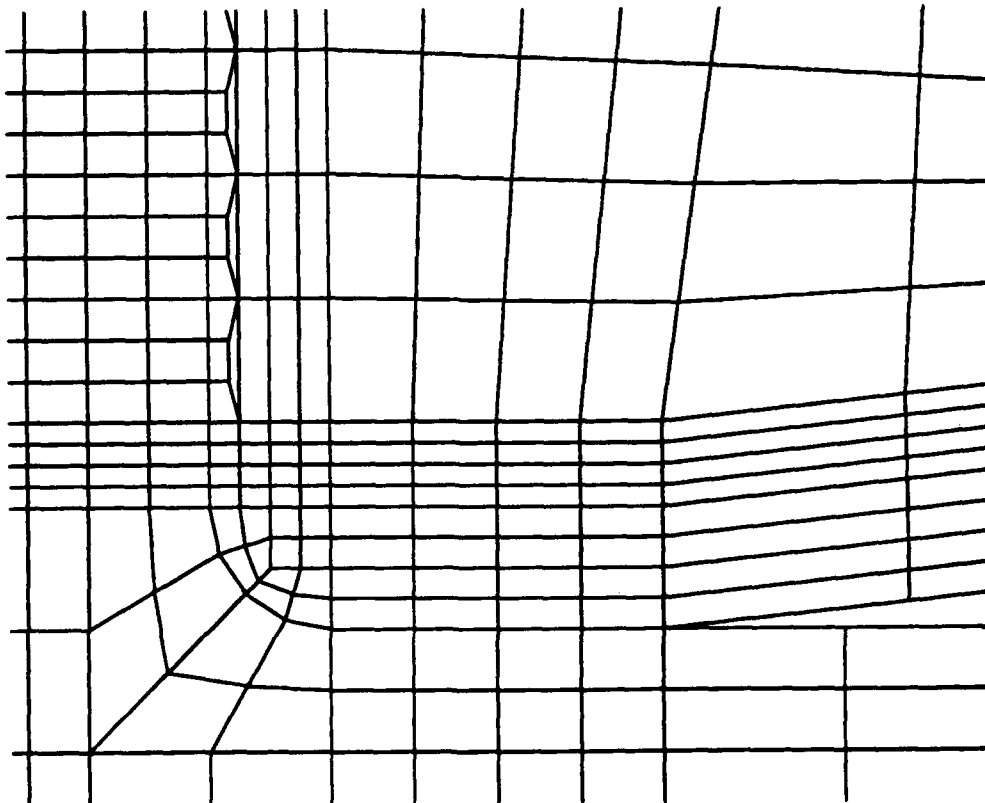
Graduate Student

John S. Strenkowski

Professor

Department of Mechanical and Aerospace Engineering

The objective of this research project is to understand the wear process of a diamond tool. Three wear mechanisms will be analyzed: the abrasive wear of diamond caused by the flow of workpiece material, chemical wear processes that remove carbon from the tool, and oxidation/graphitization of the diamond at elevated temperatures. The parameter that determines which wear mode is predominant is temperature of the chip/tool interface. Using a 2-D Eulerian model of orthogonal cutting, the chip velocity and the temperature distribution in the chip and the tool can be determined. By coupling this information with a model of wear, the effect of tool geometry on tool wear can be predicted. Ultimately changes in the tool geometry can then be used in a model to study the variation of the wear process with time and its influence on the machined surface.



8.1 INTRODUCTION

Diamond tools can wear in a variety of modes. Wilks and Wilks [1] identify fracture, mechanical attrition, thermal degradation, and chemical wear. This section concentrates on tool wear during single point diamond turning (SPDT) of copper. It is assumed that cutting conditions are chosen such that fracture of the tool can be avoided.

Both the flank face and the rake face of a tool exhibit wear during single point diamond turning. Flank face wear is characterized by the growth of a wear land on the tool flank parallel to the surface of the workpiece. This wear land is relatively smooth. The wear on the rake face, however is characterized by deep craters along the rake face of the tool and almost negligible wear near the cutting edge. If mechanisms causing wear can be identified, tool life could be extended and surface finish improved.

One good candidate for a chemical wear mechanism is solid state diffusion of carbon atoms directly into the workpiece. Diamond tools wear very quickly when machining materials in which carbon is highly soluble (e. g. steel and silicon) implying that solid state diffusion is the dominant process governing the wear rate for these materials. In addition, the governing equation for solid state diffusion has an exponential dependence on temperature. This dependence means that small changes in temperature have large effects on the wear rate, and would allow deep wear pits to occur close to areas of almost no wear (as observed by Drescher [2]). This section therefore concentrates on diffusion as the dominant chemical wear mechanism.

Temperature is the key to predicting wear in a diffusion wear model. The average temperature in SPDT has been directly measured by Iwata [3], but a discrete temperature distribution along the rake and flank face has yet to be determined by experiment. The temperature distribution can be evaluated using advanced numerical techniques, such as the finite element method. A model of orthogonal cutting has been developed by Strenkowski and Carroll [4, 5] and by Strenkowski and Moon [6] for chip geometry and temperature prediction. These models were developed to study conventional machining processes. However, they can be applied to micromachining in single point diamond turning with additional refinement. Specifically, modifications must be made to the boundaries to accurately simulate thermal boundary conditions. This is necessary because the boundaries of the cutting model are so close to the cutting zone that the temperatures within the workpiece and tool are extremely sensitive to the thermal boundary conditions. This sensitivity can be eliminated by incorporating thermal infinite elements in the cutting model.

It is found that the large thermal conductivity of diamond and its strong dependence on temperature play a critical role in the resulting wear patterns along the chip-tool interface. Near the tool edge where the temperatures are the highest, the thermal conductivity is significantly reduced, leading to more heat remaining in the chip. As a result, the greatest wear occurs at a point along the rake face which is far from the tool edge.

Two other wear mechanisms contribute to the wear of a diamond tool: mechanical attrition and thermal degradation. Mechanical attrition is the abrasive removal of atoms from the tool by the workpiece. The wear rate due to abrasion is a function of the relative hardness of the workpiece and the tool (which is a function of the crystal orientation of the tool). It is negligible when machining soft materials, but is the dominant mechanism for tool material removal for hard materials (i. e. during diamond lapping of a diamond tool). Thermal degradation occurs at elevated temperatures when the diamond crystal oxidizes and/or turns into graphite (graphitization). These reactions occur at high temperatures and are dependent on the crystal structure of the diamond, the presence of oxygen, and pressure. Diamond tools start to oxidize and graphitize at 1500 °C in an inert environment [7]. In the presence of oxygen this reaction can start to occur at 500 °C [1]. The experimental work and theoretical cutting models indicate that temperatures of this magnitude are not reached in the SPDT of copper, therefore thermal degradation is an unlikely wear mechanism.

8.2 DIAMOND TOOL WEAR EXPERIMENTS

Diamond tools have been shown to demonstrate significant wear. The wear of natural diamond tools of zero degree rake angle cutting OFHC copper has been measured experimentally by Drescher [2]. Drescher's wear experiment used a straight-edged natural diamond tool to plunge cut to a depth of 0.125 μm at a cutting speed of 600 m/min. After 8000 seconds of cutting, wide variations in the depths of wear on the rake face were determined from an interferogram. Wear near the cutting edge was not discernable, but a maximum wear depth of 3 μm occurred at a point 8-10 times the depth of cut from this edge.

In another series of experiments, the average temperature along the chip/tool interface was measured for SPDT of OFHC copper [3]. Two tools were utilized in this work including a single crystal diamond with a nose radius of 1.0 mm, and a sintered diamond with a nose radius of 0.2 mm. Both tools had a zero rake angle. For these cutting experiments, the cutting speed was 754 m/min, with a depth of cut of 50 μm and an infeed of 1 μm . The average undeformed chip thickness for the sintered diamond tool was 0.35 μm and 0.16 μm for the single crystal tool. It was found that the temperature was highly dependent on the type of diamond tool. The mean temperature for the sintered diamond tool was found to be 250 °C, while the temperature for the single crystal diamond was considerably less (110 °C).

Crompton, Hirst and Howse [8] measured the coefficient of abrasive wear of a diamond rubbing on a copper surface at low speed (88 mm/s). At this speed temperature effects will be small and all the wear will be due to abrasion. Their results are summarized in Table 1.

Experiment	Diamond Crystal Orientation	Copper Hardness (Vickers Pyramid) (GPa)	Wear Coefficient ($10^{-12} \text{ mm}^3 \text{ N}^{-1} \text{ m}^{-1}$)
Case I	[010]	0.92	4.1
Case II	[110]	0.96	2.1
Case III	[110]	0.75	1.4

Table 1: Measured Wear Coefficients (Diamond on Copper)

In one of Drescher's most recent cutting force experiments (see Section 7), OFHC copper (Vickers hardness 0.76 GPa) was machined with a round nose SPDT tool. The cutting conditions for this experiment were $V = 10 \text{ m/s}$, rake angle $= 0^\circ$, clearance angle $= 6^\circ$, depth of cut $= 4.8 \mu\text{m}$, feed $= 4.0 \mu\text{m}$, and nose radius $= 1500 \mu\text{m}$. The measured cutting force was 0.43 N, the measured thrust force was also 0.43 N, and the coefficient of friction (μ) was calculated to be 0.2. The crystal orientation of the diamond was not available so all three cases of Crompton, Hirst and Howse will be used for comparison.

8.3 DIAMOND TOOL WEAR MODELS

It is assumed that wear along the rake face is diffusion controlled. Solid state diffusion of carbon atoms from the diamond tool to the workpiece is governed by Fick's First law, which includes a diffusion coefficient with an exponential dependence on temperature. By further assuming that the carbon concentration gradient can be estimated by setting the separation between the tool and the workpiece to be equal to the atomic lattice parameter and by setting the concentration of carbon atoms in the tool to zero, the temperature (T_{req}) necessary to account for the observed wear can be calculated:

$$T_{\text{req}} = \frac{\frac{-\Delta E_D}{R}}{\log \left(\frac{d_{\text{wear}} l_{\text{dia}}}{t_{\text{mach}}} \right) - \log D_0} \quad (1)$$

where D_0 is the frequency factor, ΔE_D is the activation energy, l_{dia} is the lattice parameter of diamond, d_{wear} is the depth of wear damage, t_{mach} is the machining time, and R is the universal gas constant.

Equation (1) can be applied to Drescher's wear experiment by setting $R = 1.98 \text{ cal / mole K}$, $t_{\text{mach}} = 8000 \text{ sec}$, and $l_{\text{dia}} = 0.35 \text{ nm}$. Note that the values for the interstitial diffusion of carbon atoms into copper as given by Bergner [10] are $D_0 = 74. \times 10^{-6} \text{ m}^2 / \text{sec}$ and $\Delta E_D = 37.9 \text{ kcal / mole}$. The solution to equation (1) predicts a required temperature 200°C for negligible wear (5nm) and a temperature of 270°C for a $3 \mu\text{m}$ wear pit. These temperatures are attainable for these cutting conditions, as reported by Iwata. Therefore, solid-state diffusion of carbon from the diamond tool seems to be a plausible mechanism for rake face wear.

Wear on the flank face of the tool is characterized by the growth of a wear land parallel to the cut surface. The craters seen on the rake face are not present, implying that diffusion is not the dominant form of wear and that the temperature is constant along the tool flank. The high speed flow of workpiece material across the flank is an ideal condition for abrasive wear. Abrasive wear is governed by Archard's wear law [8]:

$$W = K_1 P S \quad (2)$$

where W is the wear, P is the load, S is the sliding distance and K_1 is the wear per unit load per unit sliding distance or coefficient of wear. This equation can be used to find the wear rates by replacing the sliding distance S by the relative sliding velocity.

The wear rates predicted by the abrasion model can be compared to those predicted by the diffusion model by examining the following particular case. Substituting Crompton's measured wear coefficients and Drescher's cutting velocity and loads into equation (2) results in predicted wear rates of 17.6, 8.60, and $6.02 \times 10^{-12} \text{ mm}^3 \text{ s}^{-1}$ for Crompton's Cases I, II, II, respectively. The wear rate predicted by the diffusion model can be found by assuming a temperature, solving Fick's law for the diffusion flux, and then multiplying by the assumed area of contact. Data from Drescher's same tool force experiment will be used. The contact area assumed was equal to the undeformed chip thickness (a reasonable approximation since the cutting force and the thrust force were equal in this experiment). The predicted wear rates are shown in Table 2.

Assumed Temperature ($^\circ\text{C}$)	Predicted Diffusion Wear Rate($10^{-12} \text{ mm}^3 \text{ s}^{-1}$)
200	0.0224
300	26.1
400	3740

Table 2: Predicted Diffusion Wear Rate

Comparison of the predicted wear rates for these two wear models illustrates the sensitivity of the diffusion wear rate to interface temperature. At low temperatures, wear is characterized by abrasion, while at high temperatures wear is predominantly due to diffusion. At very high temperatures oxidation or graphitization of the tool can also occur [7].

8.4 THERMAL MODEL

8.4.1 Heat Balance

The wear models demonstrate that temperature plays a critical role in predicting wear mechanisms. The overall temperature determines which wear modes are significant and the local temperature variation is important in diffusion controlled wear due to the exponential temperature dependence of the diffusion coefficient. The local temperature variation due to the sliding of the diamond over the workpiece can be analytically calculated by assuming that the area of contact between the two surfaces is determined by the yield pressure of the workpiece. An expression for the temperature rise in sliding contact derived by Bowden and Tabor is [11]:

$$T - T_{\text{bulk}} = \frac{\mu V \sqrt{L} \sigma_y \pi}{4 (k_{\text{tool}} + k_{\text{wp}})} \quad (3)$$

where T_{bulk} is the bulk temperature, μ is the coefficient of friction, V is the cutting velocity, L is the load (measured thrust force), σ_y is the yield stress of the softer material, and k_{tool} and k_{wp} are the thermal conductivities of the tool and workpiece. Note that for the SPDT of copper $\sigma_y = 0.31$ GPa [12], $k_{\text{tool}} = k_{\text{dia}} = 7.03$ W / cm K [13], and $k_{\text{wp}} = k_{\text{cu}} = 3.95$ W / cm K [13] (conductivity values at $T=100$ °C).

Using the same conditions as those in Drescher's tool force experiment in Equation (3) results in a $T - T_{\text{bulk}}$ value of 13.2 °C. This variation is significant in the diffusion governed wear regime. Note that this is a local temperature variation due only to sliding and the temperature rise caused by plastic work is not included in this model.

The bulk temperature can also be estimated analytically if the cutting force is known and a heat balance is performed. The heat flows out of the workpiece are shown in Figure 1.

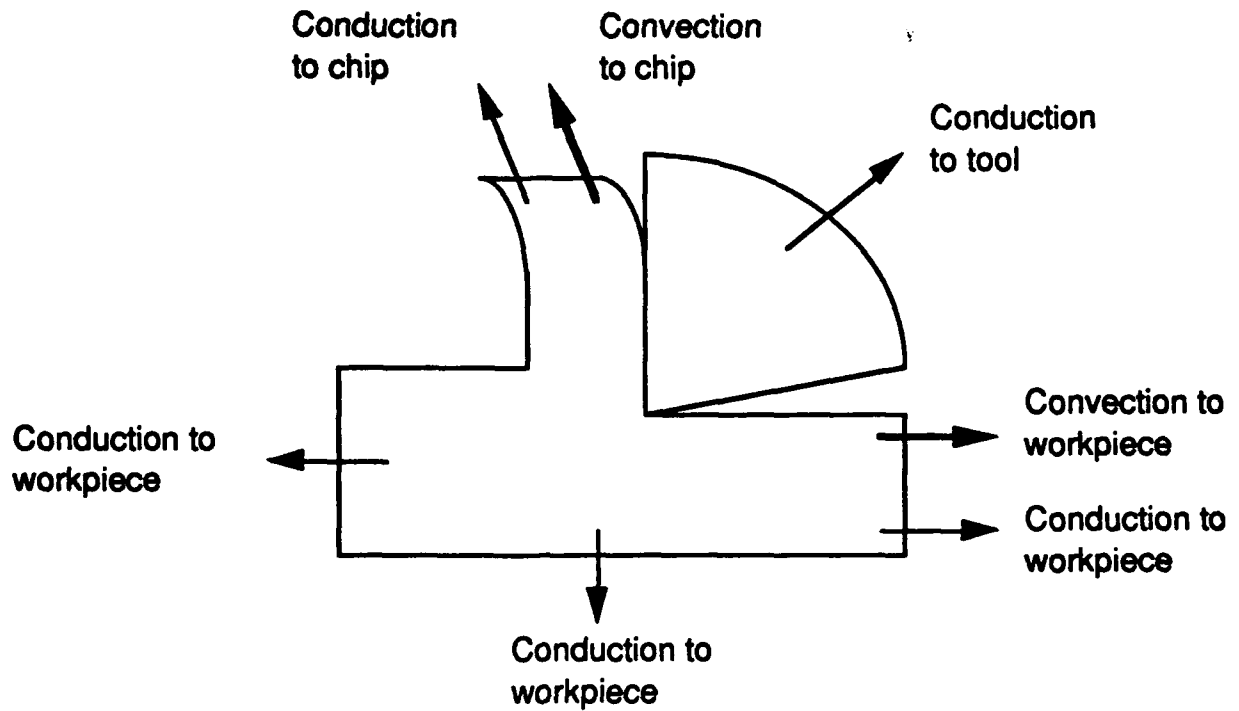


Figure 1: Heat flow from cutting model

For steady state cutting conditions the heat flow out of the model must be balanced by the heat input. Expressed mathematically, this leads to the following equation:

$$\dot{Q}_{in} = \dot{Q}_{tool} + \dot{Q}_{ch-cd} + \dot{Q}_{ch-cv} + \dot{Q}_{wp-cd} + \dot{Q}_{wp-cv} \quad (4)$$

where \dot{Q}_{in} is the heat input, \dot{Q}_{tool} is the heat flow out of the tool due to conduction, \dot{Q}_{ch-cd} is the heat flow out of the chip due to conduction, \dot{Q}_{ch-cv} is the heat flow out of the chip due to convection, \dot{Q}_{wp-cd} is the heat flow out of the workpiece due to conduction, and \dot{Q}_{wp-cv} is the heat flow out of the workpiece due to convection.

The heat flow into the workpiece can be estimated by assuming that all the energy input to the cutting process results in heat (elastic effects are negligible), which gives:

$$\dot{Q}_{in} = F_{cut} V \quad (5)$$

where F_{cut} is the measured cutting force and V is the cutting velocity. The conductive heat flow in the tool can be found by assuming annular one dimensional heat flow. This means a high

temperature region (T_{cut}) at some small radius (r_{in}) from the tip and room temperature (T_{room}) at a large radius from the tip (r_{out}). This heat flow is governed by [14]:

$$\dot{Q}_{tool} = \frac{T_{cut} - T_{room}}{\log \frac{r_{out}}{r_{in}}} \frac{\theta_{tool}}{2\pi} k_{dia} w_{cut} \quad (6)$$

Where θ_{tool} is the included angle of the tool in radians, and w_{cut} is the width of the cut. A further assumption must be made regarding the ratio between r_{out} and r_{in} , but since the logarithm of this ratio is used in the calculation its value does not change rapidly. For example, $\log(r_{out}/r_{in})$ varies from 3.9 to 5.3 for a r_{out}/r_{in} ratio of 50 to 200 respectively.

In a similar way, the conductive heat flow from the workpiece can be determined by substituting k_{cu} for k_{dia} and $\theta_{workpiece}$ for θ_{tool} . The conductive heat flow out of the chip will be included with that out of the workpiece and $\theta_{workpiece}$ will be set equal to π .

The convective heat flow out of the chip can be determined by:

$$\dot{Q}_{ch-cv} = \dot{m}_{chip} c_{p-cu} (T_{cut} - T_{room}) \quad (7)$$

where \dot{m}_{chip} is the mass flow rate of the chip, and c_{p-cu} is the specific heat of the copper (0.385 W s / gm K). The convective heat flow from the workpiece cannot be directly calculated. However, since most of the heat is generated in the shear zone (where the mass flow is directed into the chip by the tool), the convective heat flow into the workpiece should be less than that of the chip. As a result,

$$\dot{Q}_{wp-cv} = C_{cv} \dot{Q}_{ch-cv} \quad (8)$$

where C_{cv} is assumed to be constant, with a value between 0 and 1. Substituting Equations (5) through (8) into Equation(4) results in:

$$\dot{Q}_{in} = (C_{tool} + (1+C_{cv}) C_{ch-cv} + C_{wp-cd}) (T_{cut} - T_{room}) \quad (9)$$

where the subscripts of C_i use the same definitions as \dot{Q}_i . The C_i coefficients determine the relative heat flows to the various sinks.

Using values from Drescher's cutting force experiment (see Section 7), the heat transfer coefficients can be determined from Equations (6) and (7) as shown in Table 3. Note that for these conditions, $\dot{m}_{chip} = 1.72 \times 10^{-3}$ gm/s, $\theta_{tool} = 1.47$, and $w_{cut} = 120 \mu\text{m}$.

Constant	(r_{out} / r_{in})	Value (10^{-3} W / K)
C_{ch-cv}	-	0.663
C_{tool}	50	3.72
C_{tool}	100	4.27
C_{tool}	200	5.03
C_{wp-cd}	50	4.47
C_{wp-cd}	100	5.14
C_{wp-cd}	200	6.06

Table 3: Heat Flow Coefficients

Note in Table 3 that the chip convective heat flow is about an order of magnitude less than the convective heat flow of either the tool or the workpiece, and that the convective heat flows of the tool and the workpiece are approximately equal. By assuming a ratio for (r_{out} / r_{in}) and $\dot{Q}_{wp-cv} / \dot{Q}_{ch-cv}$ the temperature rise can be calculated from Equation (9). Table 4 demonstrates that the solution to Equation (9) is not overly sensitive to the assumptions that must be made to solve the problem analytically. That is, changing the ratio of (r_{out} / r_{in}) from 50 to 200 and the value of $\dot{Q}_{wp-cv} / \dot{Q}_{ch-cv}$ from 0 to 1, only changes the temperature rise from 346 to 452 °C.

$\dot{Q}_{wp-cv} / \dot{Q}_{ch-cv}$	(r_{out} / r_{in})	$T_{cut} - T_{room} (^\circ\text{C})$
1	50	365
1	100	426
1	200	486
0	50	346
0	100	400
0	200	452

Table 4: Calculated Temperature Rise

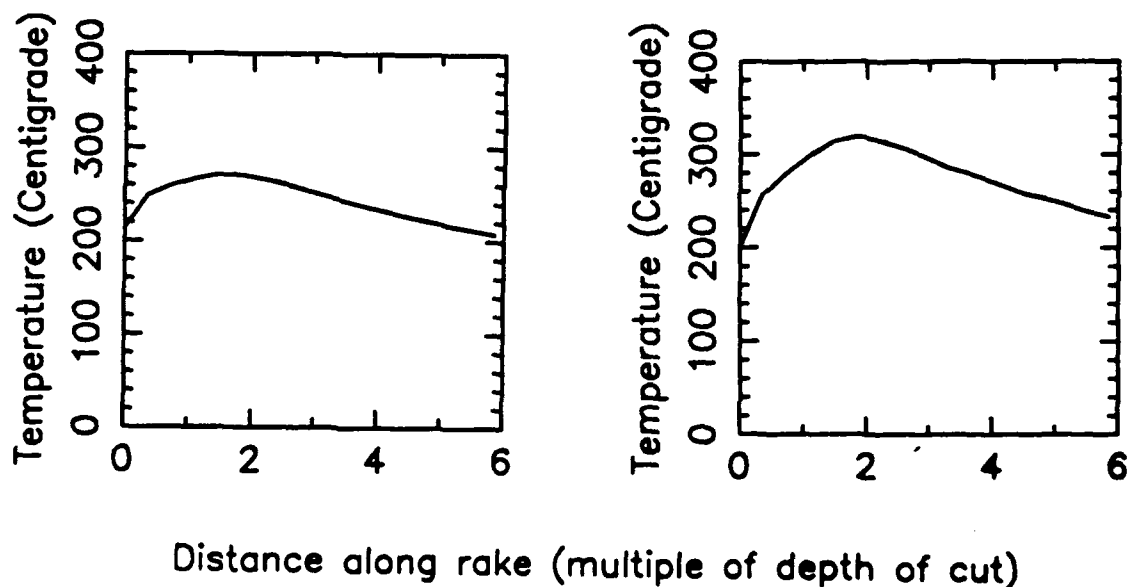
The calculated temperatures are well below the oxidation temperature of diamond, but they are high enough so that diffusion wear will predominate over abrasive wear as the primary wear mechanism.

8.4.2 Finite Element Model of Rake Face

Using the Eulerian cutting model, it is also possible to determine the temperature and wear pattern on the rake face. The tool and the workpiece were modeled with 310 and 882 linear isoparametric elements, respectively. Thermal infinite elements were used along the interior tool boundary and the lower edge of the workpiece. The remaining edges of the workpiece were treated as insulated. Note that the infinite elements used in the thermal model have become an increasingly used and

accepted means for the solution of unbounded field problems. These infinite elements complement finite elements for unbounded problems with a minimum of coding changes and without destroying the sparsity pattern of the system matrix, a major defect of other alternative techniques.

The temperature along the rake face predicted by the finite element model of Iwata's [3] cutting conditions is shown in the Figure 2(a). Note that the peak temperature of 270 °C occurs for a depth of cut of 50 μm . It was found that this depth of cut was necessary to achieve the temperature rise required for the observed rake face wear depths. It is interesting to note that the thermal conductivity of diamond is highly sensitive to temperature. The above profile was generated using the room temperature value of Type I diamond conductivity (9 W / cm K). Type I diamond was chosen due to its predominance in the natural diamond supply. Figure 3 shows the thermal conductivity of diamond as a function of temperature [15]. For the range of interest of 200-300 °C, the conductivity decreases by approximately 67% from its ambient value. When the temperature-dependent conductivity is included in the cutting model simulation, the maximum shifts even farther from the cutting edge (Figure 2(b)) as compared to the constant thermal conductivity. In addition, the maximum temperature increases to nearly 320 °C. As the conductivity decreases with the temperature near the cutting edge, more of the heat generated in the shear zone remains in the chip. This process continues until a maximum temperature is reached at some point along the rake face where the thermal conductivity of the diamond increases sufficiently.



(a) constant diamond conductivity (b) variable diamond conductivity

Figure 2 : Variation of rake face temperature along rake face

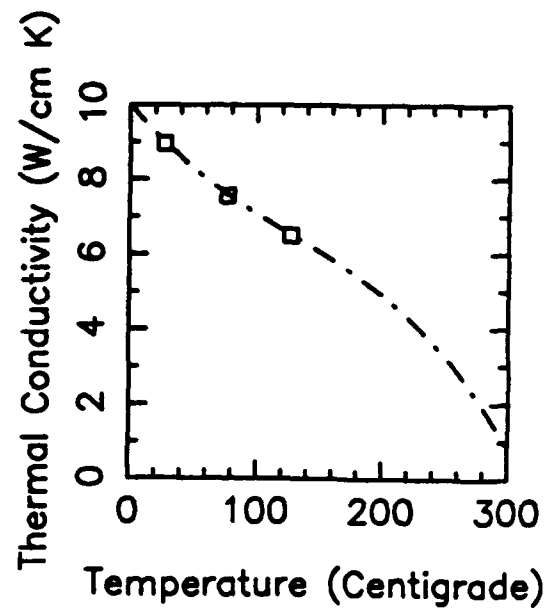


Figure 3: Type I diamond conductivity vs temperature

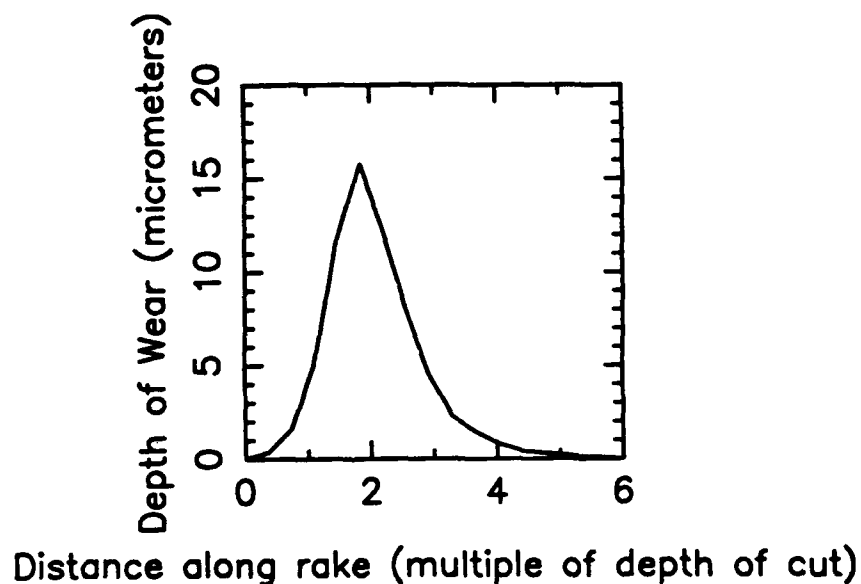


Figure 4: Predicted depth of rake face wear along rake face

Because wear is highly dependent on the temperature, the wear pattern will be similar to the temperature distribution. This can be seen in Figure 4, which shows the predicted wear depth as a function of distance along the rake face. This pattern is consistent with the observed rake face wear of Drescher's straight edge tool experiments [2]. The wear is very sudden, occurring at a point located several depths of cut from the cutting edge.

8.4.3 Finite Element Model of Rake and Flank Faces

To study the flank face of the tool, the cutting model was extended to include both a tool edge radius (B-C) and a wear land on the flank face (A-B) as shown in Figure 5. The temperature output of this model can then be used to predict diffusion wear rates.

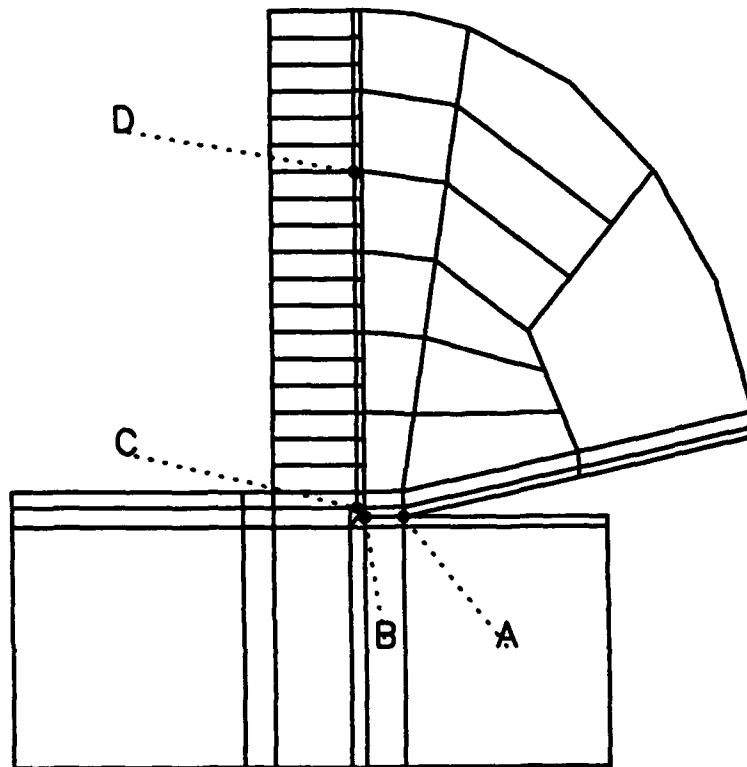


Figure 5: Extended finite element model

The cutting conditions are described in Section 7 of this report. A straight edge tool with a rake angle of 0° , and a clearance angle of 6° was used to plunge cut to a depth of $1.0 \mu\text{m}$. The cutting velocity was 10 m/s , and the measured chip thickness was $1.2 \mu\text{m}$. The thrust force and cutting force were 4.5 N and 2.6 N , respectively. After 4 km of wear, based on the model described in Section 7, the tool edge radius was $.59 \mu\text{m}$, the wear land was $1.62 \mu\text{m}$, and the flow stress was 1.89 GPa . The temperature distribution is shown in Figure 6. The resulting temperature profile along the tool edge is shown in Figure 7. Note that the maximum temperature occurs at about 5 times the depth of cut along the rake face of the tool and that the temperature along the tool nose (B-C) averages about 170°C . The rake face peak temperature location is consistent with previous results in that it occurs well up the rake face of the tool (where wear craters were observed experimentally).

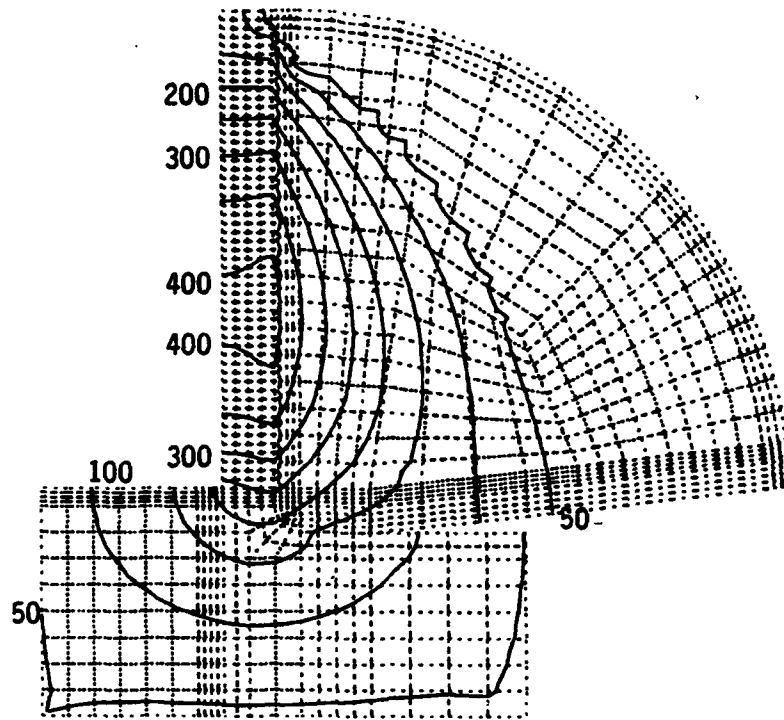


Figure 6: Predicted temperature contours (°C)

The predicted temperature distribution can then be used to predict the wear rate due to diffusion. This calculation yields a predicted wear rate of $881 \cdot 10^{-12} \text{ mm}^3 / \text{s}$. The temperature can also be predicted from the bulk tool model by assuming a value of $(r_{\text{out}} / r_{\text{in}})$ of 100, and setting the ratio of $\dot{Q}_{\text{wp-cv}} / \dot{Q}_{\text{ch-cv}}$ to 1. This value is reasonable since the depth of cut was only slightly greater than the tool edge radius. Recalculating the diffusion wear rate calculation with the temperature rise predicted by the bulk tool model ($\Delta T = 176 \text{ }^\circ\text{C}$) results in a predicted wear rate of $1460 \cdot 10^{-12} \text{ mm}^3 / \text{s}$. The abrasion model for this case predicts a wear rate from 63 to $184 \cdot 10^{-12} \text{ mm}^3 / \text{s}$, depending on wear coefficient. These values can be directly compared to Drescher's experiment by noting that at 2 km of wear the edge radius was $0.18 \text{ } \mu\text{m}$ and the wear land was $1.14 \text{ } \mu\text{m}$. The edge radius and wear land present at 4 km result when 0.233 mm^3 of tool material has been worn away. The resulting wear rate is $1160 \cdot 10^{-12} \text{ mm}^3 / \text{s}$. These results are summarized in Table 5.

Model	Predicted Wear Rate ($10^{-12} \text{ mm}^3 / \text{s}$)
Abrasion (Crompton's average wear coefficient)	114
Diffusion (Finite element model temperature)	881
Diffusion (Bulk tool model temperature)	1460
Drescher's Experiment	1160

Table 5: Comparison of Wear Rates

For these cutting conditions diffusion is the dominant wear mode. Note that the linear growth rate of the wear land observed by Drescher is consistent with both the abrasion and diffusion models. In the abrasion model, the linear increase in load (thrust force) results in a linearly increasing wear rate, but the volume worn away by increases in the length of the wear land also increases linearly, making the wear land length to cutting distance curve a straight line. The diffusion model predicts wear rate increasing linearly with contact area, which is also consistent with linear growth of the wear land.

The cutting model can be further validated by comparing predicted tool forces to Drescher's measured values. The model is in agreement with the thrust force measurement (4.3N predicted vs. 4.5 N measured), but predicts about 5 times the measured cutting force. The predicted normal pressure along the tool edge is shown in Figure 8. The variation in pressure shown occurs about at the undeformed chip thickness up the rake face of the tool and may be a factor in the poor cutting force prediction of the model.

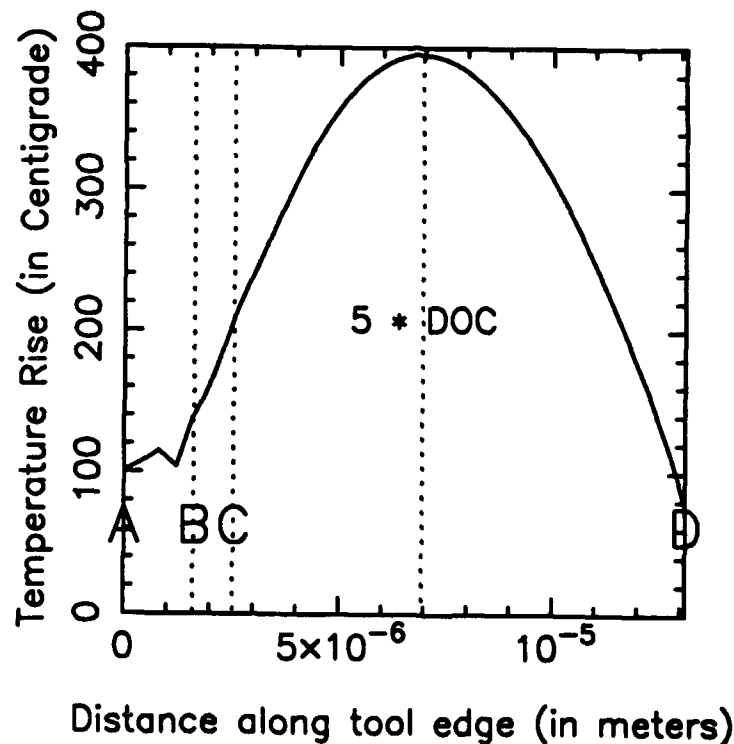


Figure 7: Variation of temperature along tool edge

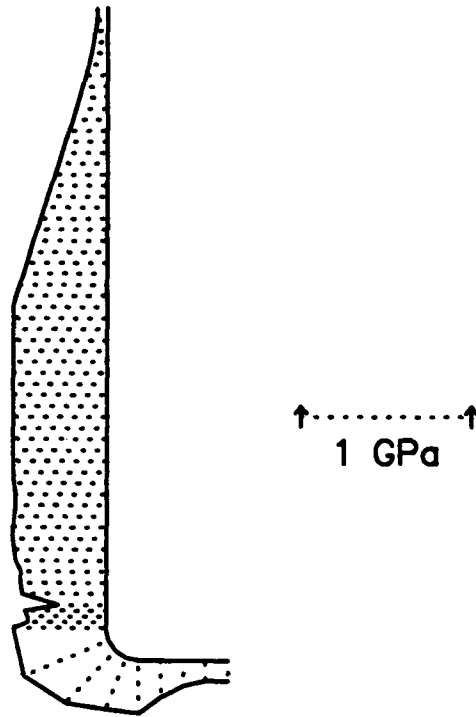


Figure 8: Pressure variation along tool edge

8.5 CONCLUSIONS

Abrasion and diffusion wear mechanisms contribute to diamond tool wear. The oxidation/graphitization wear mechanism does not occur because the temperatures never reach the required elevated values. On the rake face, the dominant wear mechanism is diffusion. The governing equation for diffusion predicts extreme sensitivity of wear to the temperature distribution in the tool. In the case of copper a change of only 6 °C results in a doubling of the diffusion coefficient, which also doubles the wear rate. These wear rates are consistent with experimentally observed temperatures, as well as wear patterns predicted with a finite element cutting model. Maximum temperature and the maximum wear occurs at a point located several depths of cut from the cutting edge of the tool. It was found that the temperature dependence of the thermal conductivity of the diamond caused the maximum temperature to not only shift away from the cutting edge, but also increase in magnitude.

Abrasive wear on the flank face (where the temperature is lower and the material flow velocity is higher) may be significant depending on the temperature of the interface. A heat transfer analysis based on a simple analytical model showed that diffusion may also be the dominant wear mechanism on the flank face. Oxidation and graphitization temperatures cannot be generated, so this wear mechanism is probably unimportant in the diamond turning of copper.

Future work will revolve around further improving the cutting model. The thermal model will be modified to include imperfect contact between the chip and the rake face of the tool and experiments will be conducted to verify the wear models.

References

1. Wilks, J. and E. Wilks, Properties and Applications of Diamond, 1991, Oxford: Butterworth-Heinemann Ltd.
2. Drescher, J. and T.A. Dow, Surface Finish, "Tool Force Relationships in Diamond Turned Copper", in *Precision Engineering Annual Report*, 1990, North Carolina State University: Raleigh, NC.
3. Iwata, K., T. Moriwaki, and K. Okuda, "A Study of Cutting Temperature in Ultra-High Precision Diamond Cutting of Copper". 15TH North American Manufacturing Research Conference Proceedings, 1987 SME *Manufacturing Technology Review*, 1987. 2: p. 510-515.
4. Carroll III, J.T. and J.S. Strenkowski, "Finite Element Models of Orthogonal Cutting With Application to Single Point Diamond Turning", *International Journal of Mechanical Sciences*, 1988. 30: p. 899-920.
5. Strenkowski, J.S. and J.T. Carroll III, "An Orthogonal Metal Cutting Model Based on an Eulerian Finite Element Method. in Manufacturing Processes", *Machines and Systems, Proceedings of the 13th NSF Conference on Production Research and Technology*, . 1986.
6. Strenkowski, J.S. and K.J. Moon, "Finite Element Prediction of Chip Geometry and Tool/Workpiece Temperature Distributions in Orthogonal Cutting", *ASME Journal of Engineering for Industry*, 1990. 112: p. 313-318.
7. Davies, G. and T. Evans, "Graphitization of diamond at zero pressure and at a high pressure", *Proc. R. Soc. Lond. A.*, 1972. 328: p. 413-427.

8. Crompton, D., W. Hirst, and M.G.W. Howse, "The wear of diamond", *Proc. R. Soc. Lond. A.*, 1973. 333: p. 435-454.
9. Bergner, D., Diffusion of Heavy Interstitials in Metals. Metallurgia I Odlewnictwo, 1986. 12(4): p. 379-400.
10. Bowden, F.P. and D. Tabor, "The Friction and Lubrication of Solids", *The International Series of Monographs on Physics*, ed. R.H. Fowler, *et al.* 1950, London: Oxford University Press.
11. *Engineering Alloys*. Sixth ed. R.C. Gibbons. 1979, American Society for Metals.
12. CRC Handbook of Chemistry and Physics. 68th ed. ed. R.C. Weast, M.J. Astle, and W.H. Beyer. 1987, Boca Raton, Florida: CRC Press, Inc.
13. Hildebrand, F.F., Advanced Calculus for Applications. 1976, Prentice-Hall Inc.
14. Journal of Physical and Chemical Reference Data. ed. D.R. Lide. Vol. 1. 1972, American Chemical Society and the American Institute of Physics for the National Bureau of Standards.

9 THE INTERACTION OF MACHINING PARAMETERS ON THE FRACTURE OF BRITTLE MATERIALS DURING MACHINING

Gary D. Hiatt

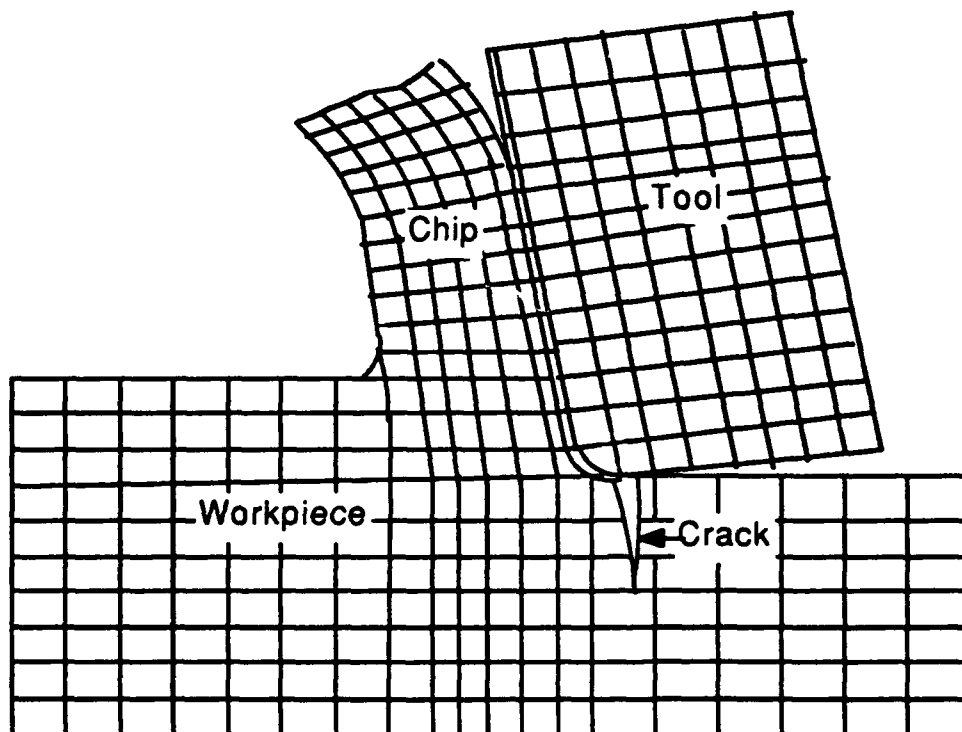
Graduate Student

John S. Strenkowski

Professor

Department of Mechanical and Aerospace Engineering

Experimental studies have shown that there exists a critical chip thickness for the transition from brittle to ductile material removal. To consistently obtain a fracture-free workpiece surface, it is necessary to understand how different material properties and machine parameters interact to influence crack initiation and propagation. The Eulerian finite element method is one way in which this understanding can be obtained. This section describes how changes in the operating conditions such as cutting speed, tool edge radius, depth of cut, and rake angle affect fracture during machining as well as offering experimental justification for the 2-D model used in the simulation.



9.1 INTRODUCTION

A technique has been presented for predicting the critical-depth parameter for SPDT of brittle materials such as silicon and germanium [1]. Using an Eulerian finite element model of orthogonal cutting, the stresses in the workpiece subsurface arising from machining can be determined as a function of the cutting conditions and tool geometry. By applying the principal of superposition in fracture mechanics to the computed stress field, the critical depth of cut below which no fracture will occur can be predicted for that particular set of machining conditions.

The computer model is a 2-dimensional plane strain simplification of the actual cutting process. Model parameters which can be varied to simulate actual machining parameters are shown in Figure 1.

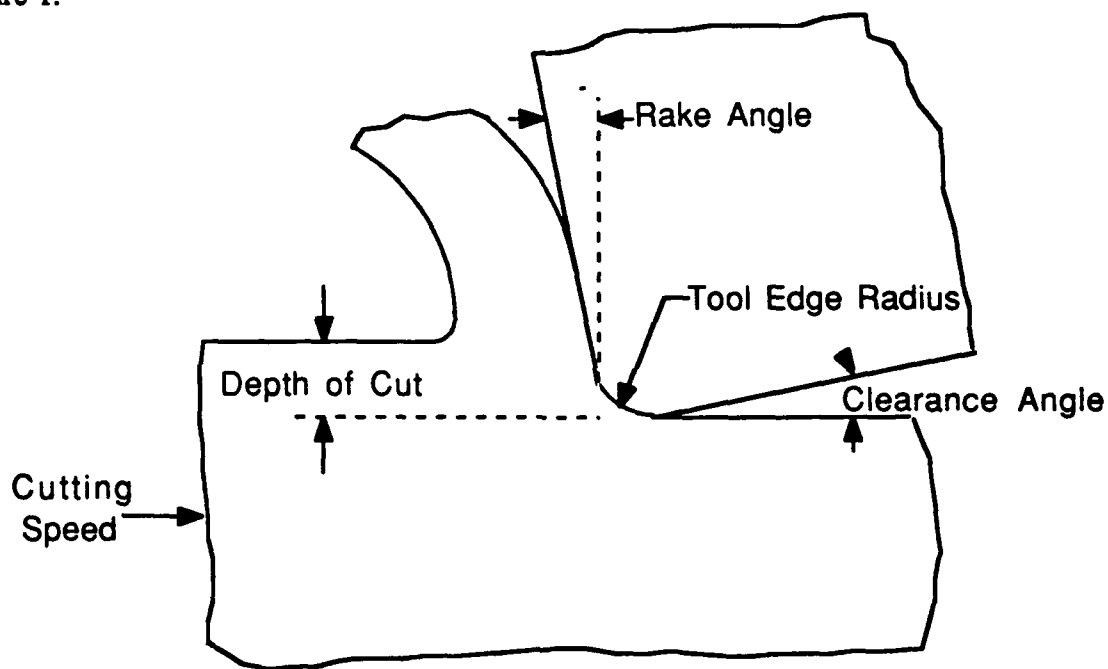


Figure 1: Machining parameters which can be varied in the 2-dimensional cutting model.

9.2 EFFECT OF CUTTING SPEED ON FRACTURE

Figure 2 shows a comparison of the steady-state stresses in the cutting direction behind the tool versus depth below the finished workpiece for the machining of germanium at three different speeds. The stresses are taken from the first column of elements behind the tool, the same column that has the crack as shown in Figure 3. The rake angle used in the simulation was -10° . All three curves show that the tensile stress is greatest at the top surface of the workpiece, and then the stress declines rapidly and becomes compressive deeper into the workpiece. This suggests that if fracture occurs, it will start at the top surface. All three curves are fairly parallel. This shows that the stress in the cutting direction decreases slightly with increases in cutting speed.

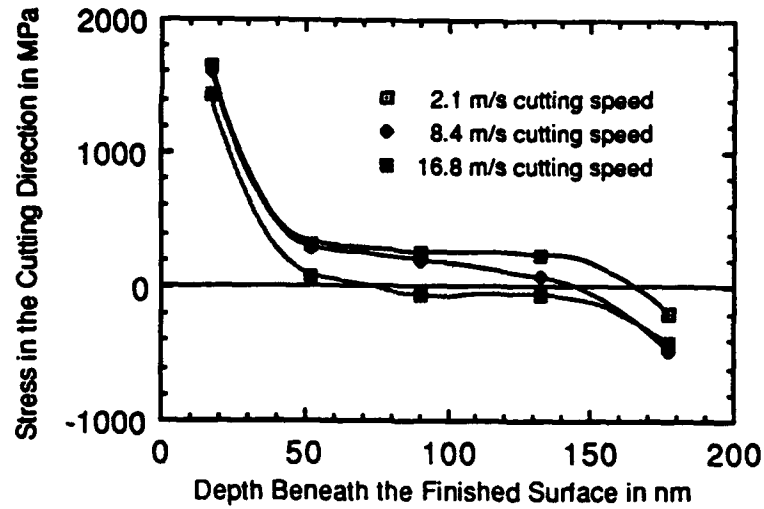


Figure 2: Comparison of stress in the cutting direction for varying cutting speeds (Germanium, -10° rake angle, 102 nm depth of cut)

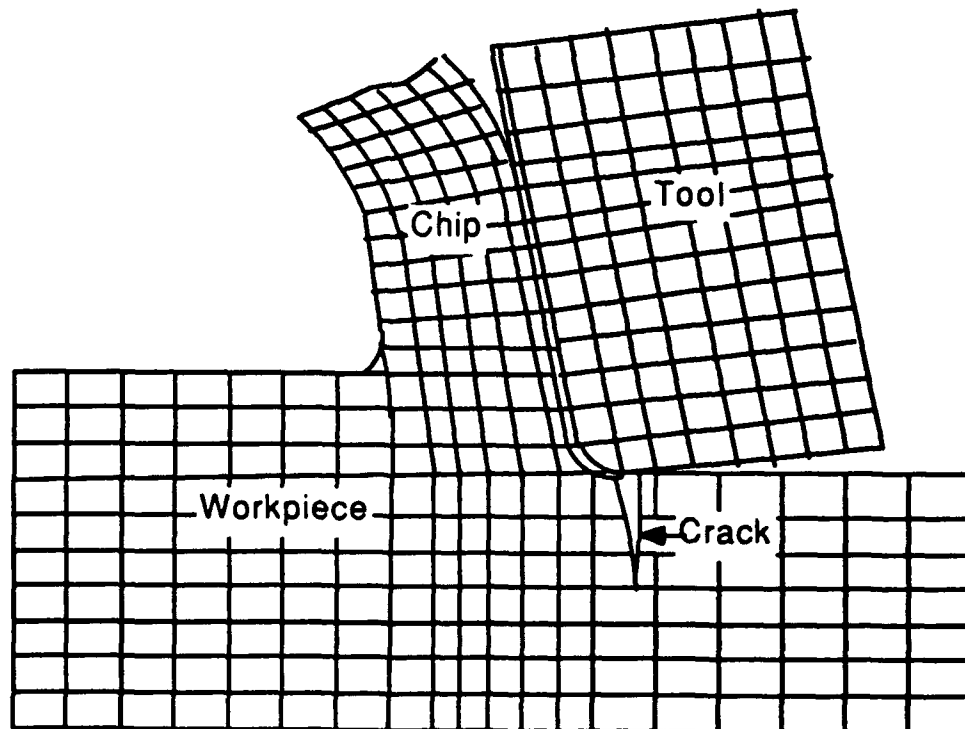


Figure 3: The steady-state stresses are taken from the first column of elements behind the tool, the same column that has the crack as shown above.

Figure 4 shows the cutting and thrust forces versus depth of cut for 2.1, 8.4, and 16.8 m/s cutting speeds. The cutting force is higher than the corresponding thrust force for each speed. The curves are also fairly parallel to one another. The cutting and thrust force curves for the 8.4 m/s speed are above those for the 2.1 m/s speed. The cutting and thrust forces for the 16.8 m/s cutting speed are above the forces for the other two speeds at the lowest depth of cut, but the 16.8 m/s forces decrease in comparison with the 8.4 m/s forces at the higher depths of cut, due to the increase in temperature with the increase in cutting speed and depth of cut.

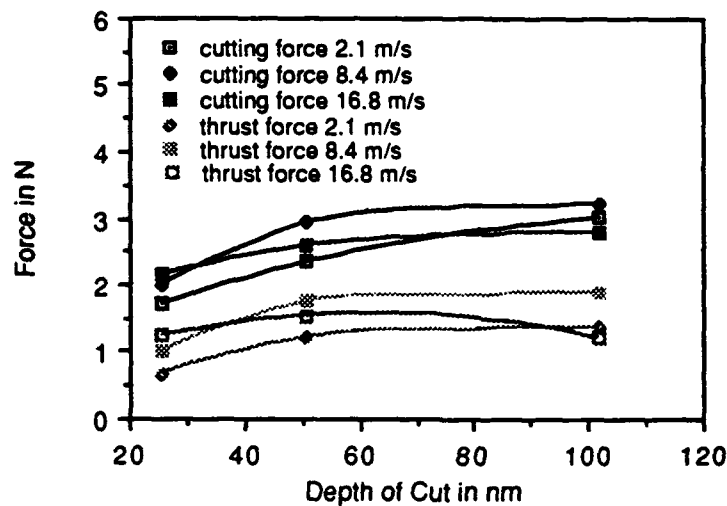


Figure 4: Comparison of machining forces for germanium with a -10° rake angle.

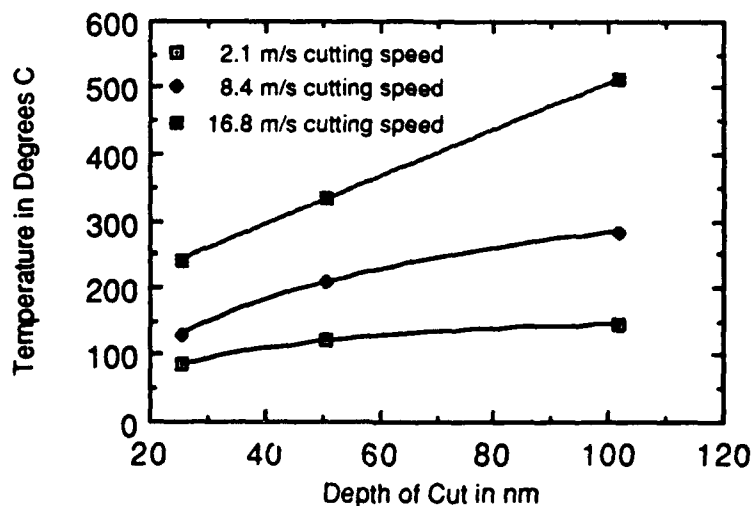


Figure 5: Predicted temperatures in the shear plane for germanium at different cutting speeds.

Figure 5 shows a comparison of the temperatures in the primary shear plane for germanium versus depth of cut at different cutting speeds. As expected, the temperature increases with an increase in depth of cut, and it increases with increasing cutting speed.

Figure 6 shows a comparison of the stress intensity factors versus crack length for three different depths of cut for the machining of germanium with a -10° rake, 6° clearance angle tool at 2.1 m/s cutting speed. The graph shows that the 51 nm depth of cut and smaller cuts should be fracture free while the 76 nm depth of cut should produce fracture. Figure 7 shows a similar comparison of the stress intensity factors for germanium with the same machining parameters except that the speed was increased to 8.4 m/s. The graphs look very similar to those in Figure 6 with the individual depth of cut curves peaking at almost the same numbers. Again ductile machining is predicted for the 51 nm depths of cut and below, while brittle machining is predicted for the 102 nm depth of cut and above. Figure 8 shows the stress intensity factors for germanium with the same machining parameters as before except that the cutting speed is increased to 16.8 m/s. The individual curves for the 25 nm and 51 nm depths of cut display similar stress intensity factors as for the lower speeds, but the 102 nm depth of cut curve peaks at a slightly lower value suggesting the higher temperatures encountered at this higher cutting speed are beginning to have an effect on the fracture of the workpiece. All three cutting speeds give the same critical depth of cut of 51 nm for the -10° rake angle. However, these results suggest that there may be a critical cutting speed above 16.8 m/s that would inhibit fracture during steady-state machining at depths of cut greater than 51 nm.

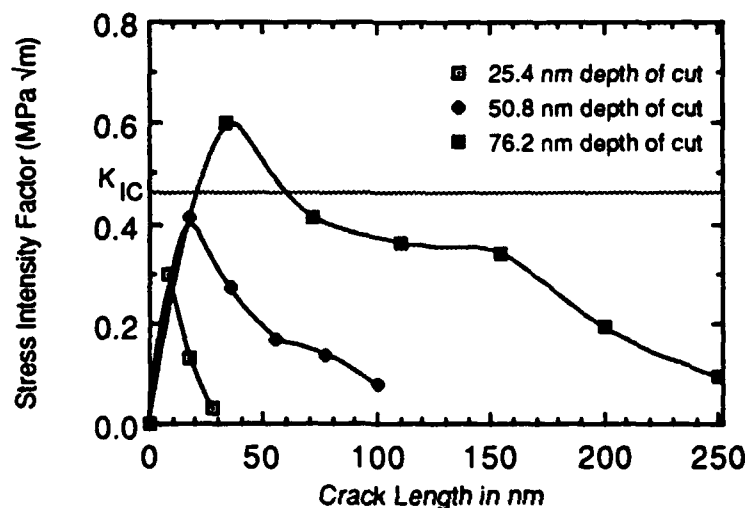


Figure 6: Stress intensity factors for the machining of germanium at various depths of cut (-10° rake, 2.1 m/s cutting speed).

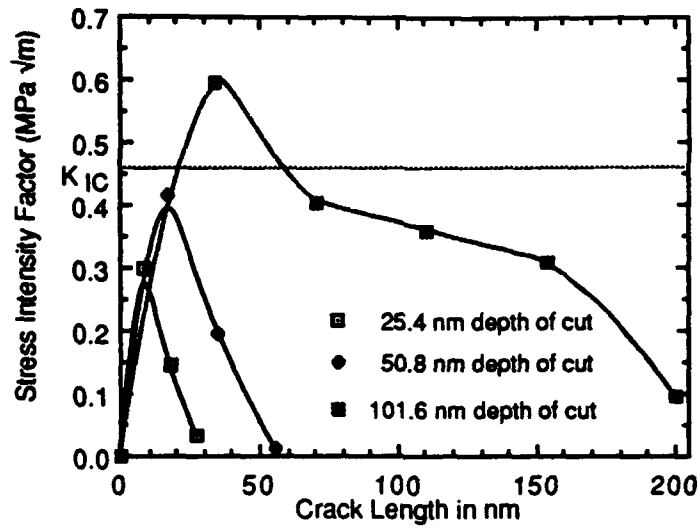


Figure 7: Stress intensity factors for the machining of germanium at various depths of cut (-10° rake, 8.4 m/s cutting speed).

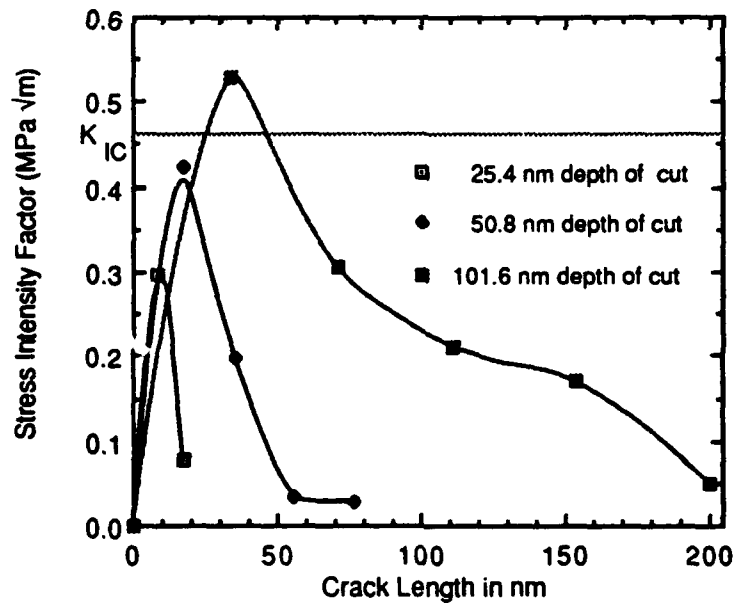


Figure 8: Stress intensity factors for the machining of germanium at various depths of cut (-10° rake, 16.8 m/s cutting speed).

9.3 THE EFFECT OF THE CUTTING EDGE RADIUS ON FRACTURE

The effect of the cutting edge radius on the quality of the finished workpiece becomes significant when the edge radius becomes comparable to the depth of cut. The size of the cutting edge radius should have a significant effect on the plastic formation of the chip, the surface deformation, and the residual stresses. As shown in Figure 9, when the depth of cut and cutting radius are comparable, then the result would be equivalent to the rake angle becoming more negative.

Various estimates of the cutting edge radius for sharp diamond tools have been reported in the literature. Asai [2] reports the edge radius from 20 to 50 nm, Evans [3] from 30 to 60 nm, Asai and Kobayashi [4] from 40 to 60 nm, Asai and Taguchi [5] from 20 to 45 nm, and Miyamoto [6] as 70 nm. Miyamoto's estimate of 70 nm was used in these simulations because the sharpest estimates are for the best case, and it was thought that a slightly worn tool edge would be closer to this value.

Three different sizes for the edge radius were used to study the effect of the cutting edge radius on fracture during machining. These were 35 nm, 70 nm, and 140 nm. Figure 10 shows the steady-state cutting stress plotted as a function of depth beneath the surface of the finished workpiece for the different cutting edge radii. The graph shows that the stress becomes less tensile as the edge radius becomes larger. This should be beneficial in reducing fracture since the stress intensity would be reduced. This graph is similar to that shown in Figure 11 which compares steady-state stress versus depth below the workpiece for different rake angles. This figure shows that the tensile stresses do not extend as far into the workpiece and are smaller in magnitude for the more negative rake angles. Thus the effect on the stress in the cutting direction of increasing the cutting edge radius is similar to decreasing the rake angle.

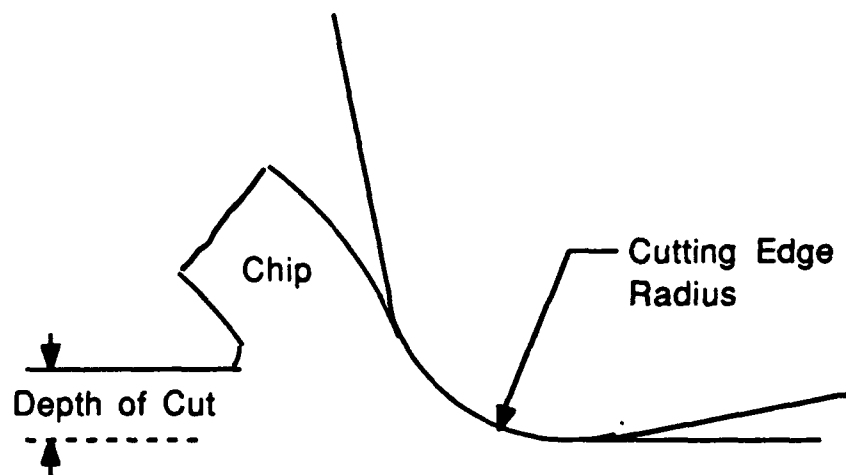


Figure 9: The nominal rake angle becomes meaningless at small depths of cut.

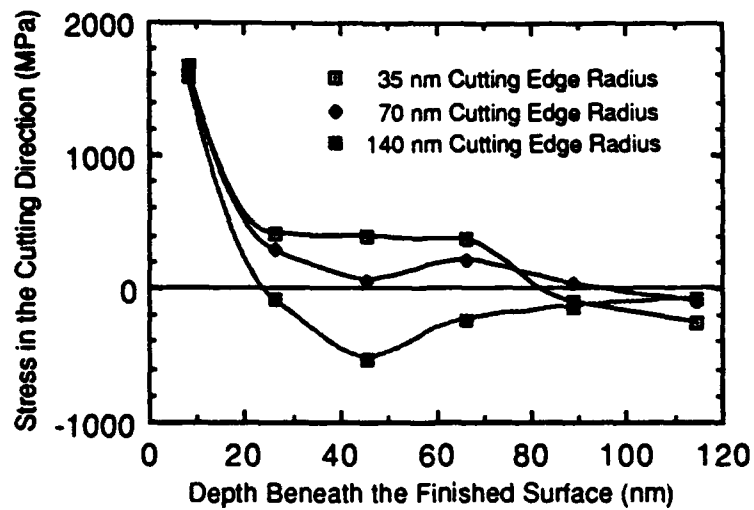


Figure 10: The steady-state stress behind the tool for various cutting edge radii (germanium, -10° rake, 2.1 m/s cutting speed, 51 nm depth of cut).

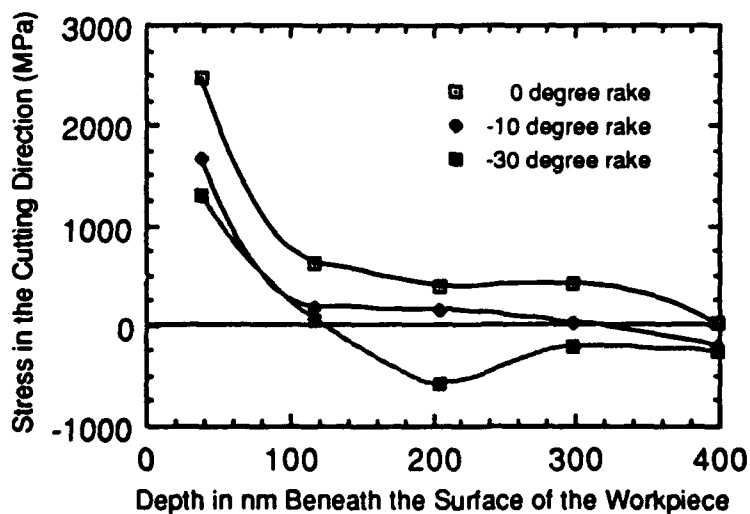


Figure 11: The variation of stress in the cutting direction with depth into the workpiece for three rake angles.

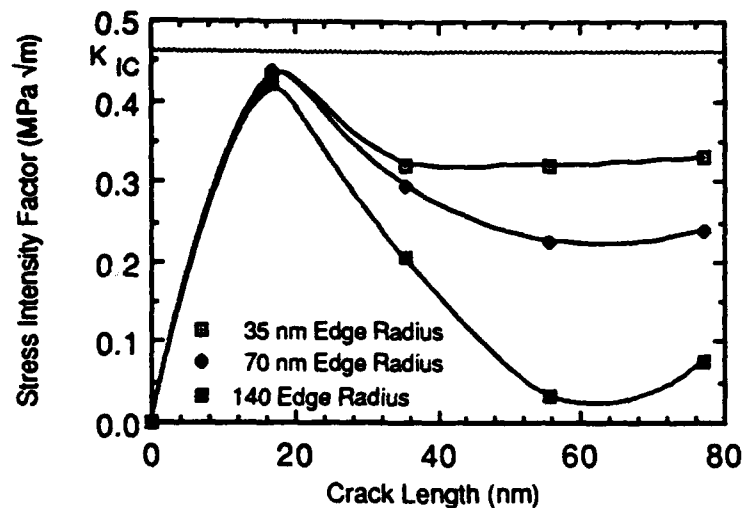


Figure 12: Comparison of the stress intensity factors versus crack length for various cutting edge radii (germanium, -10° rake, 2.1 m/s cutting speed, 51 nm depth of cut).

Figure 12 shows a comparison of the stress intensity factors versus crack length for the various cutting edge radii. The highest value of the stress intensity factor reached for all three radii is approximately $0.42 \text{ MPa}/\sqrt{\text{m}}$, which is below the fracture toughness. However Figure 12 shows that the stress intensity factors remain higher for the smaller edge radii after reaching their peaks.

Figure 13 shows a comparison of the cutting forces for the various cutting edge radii. It is seen that the cutting force increases with increasing edge radius while the thrust force increases only very slightly. The result for the thrust force was unexpected. It could be that since the model cannot account for the elastic rebound, the simulated thrust force does not include the elastic effect. However, similar results have been seen in the diamond turning of single crystal copper at small depths of cut [7]. These results are not to be confused with Drescher's model (see section 7). Drescher's wear model includes a clearance face wear land whose length increases with wear as well as a cutting edge radius which increases with wear. The finite element model does not include a wear land which should cause the thrust forces to be underestimated. Also Drescher's model was for the machining of copper, a soft ductile material while the finite element model was simulating the machining of germanium, a hard brittle material which should behave differently from copper.

Figure 14 shows the average shear plane temperatures for the different cutting edge radii. The graph shows that temperature increases as the edge radius increases. This is to be expected since most of the work done in machining is converted to heat, and increasing the edge radius would increase the work in material removal.

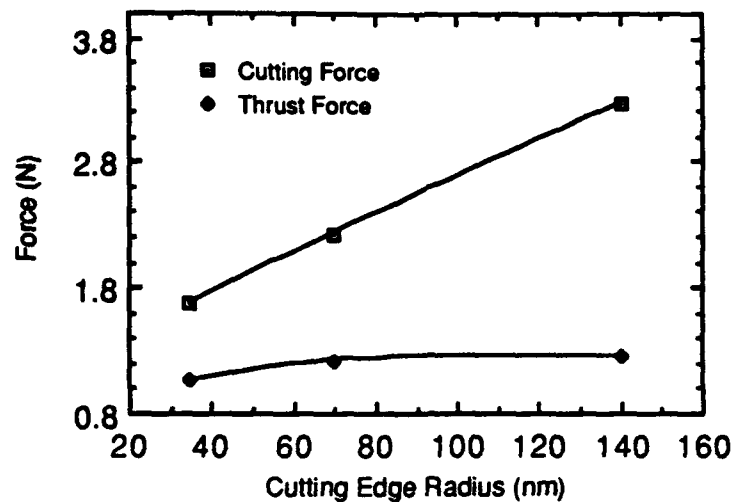


Figure 13: Comparison of cutting forces for various cutting edge radii (germanium, -10° rake, 2.1 m/s cutting speed, 51 nm depth of cut, 2.45 mm width of cut).

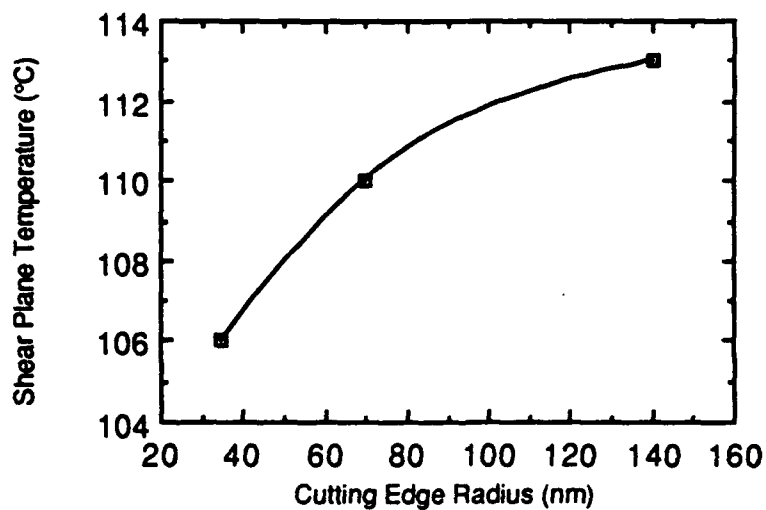


Figure 14: The average shear plane temperature for various cutting edge radii (germanium, -10° rake, 2.1 m/s cutting speed, 51 nm depth of cut).

Figure 15 shows the ratio of the thrust force to the cutting force. The ratio goes up with decreasing edge radius. This shows that the resultant force vector for the tool undergoes a rotation as the cutting edge radius changes.

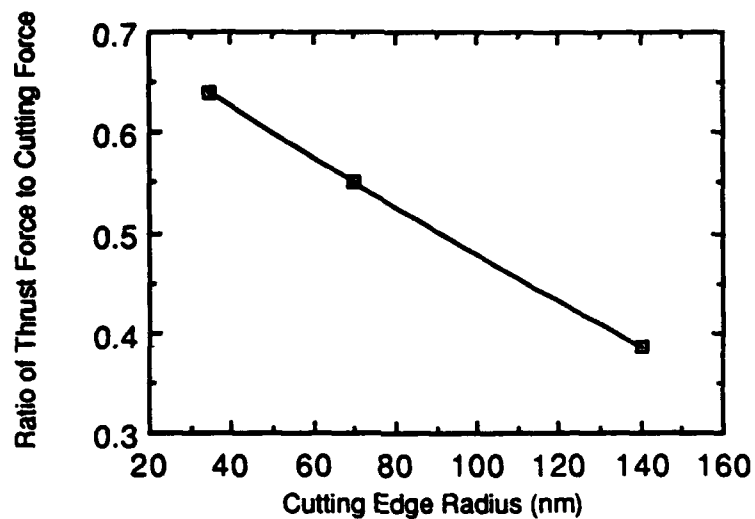


Figure 15: Effect of the cutting edge radius on the ratio of machining forces.

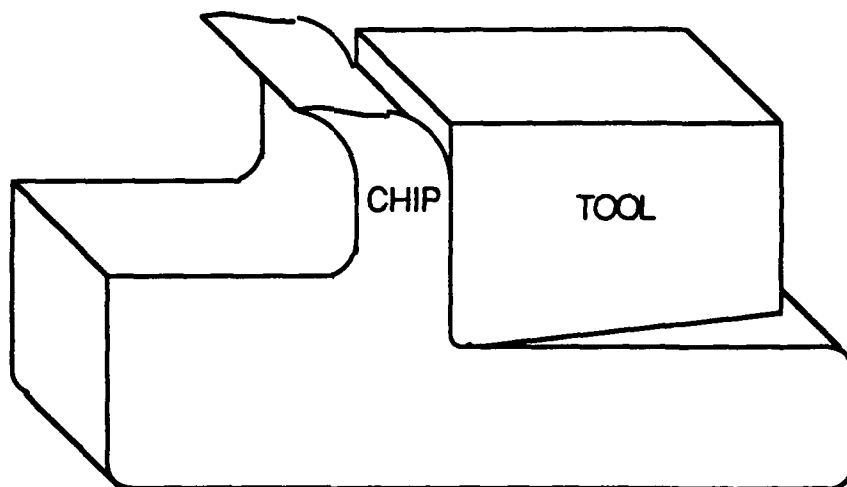


Figure 16: Idealized Chip Geometry in the Computer Model

9.4 EXPERIMENTAL RESULTS

The computer model is a 2-dimensional plane strain simplification of the actual cutting process. The tool and chip are assumed to have a constant width which is used in the tool force calculations as shown in Figure 16. Figure 17 shows a more realistic representation of the cutting geometry. The physical tool has a rounded nose which produces a chip with a varying thickness.

Most of the previous cutting experiments where force was measured used a round nosed tool as shown in Figure 17. In this case the chip changes shape. However for the model described in this section, a constant depth of cut as shown in Figure 16 was assumed. Experiments measuring tool forces and surface damage for this geometry were needed for confirmation of the finite element model.

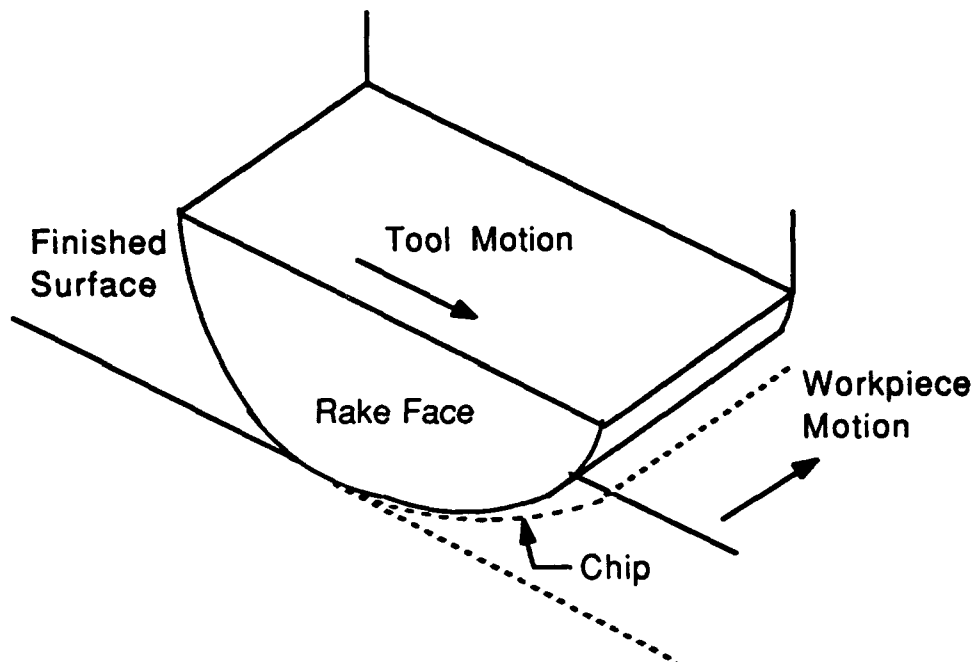


Figure 17: Approximation of Actual Chip Geometry.

9.4.1 Surface Damage

In order to more closely match the computer model and verify its predictions, a straight edge, -10 degree rake angle, 6 degree clearance angle diamond tool was chosen to machine germanium. For this rake angle, the model predicts a critical depth of cut of 50 nm as shown in Figure 18. Five depths of cut were machined in the germanium at 25, 38, 50, 64, and 76 nm. Figure 19 shows the resulting machined surfaces for these depths of cut. The surfaces for the 25 and 38 nm depths of cut show little pitting damage. The pitting increases greatly for the 50 nm depth of cut and increases greater still for each larger depth of cut.

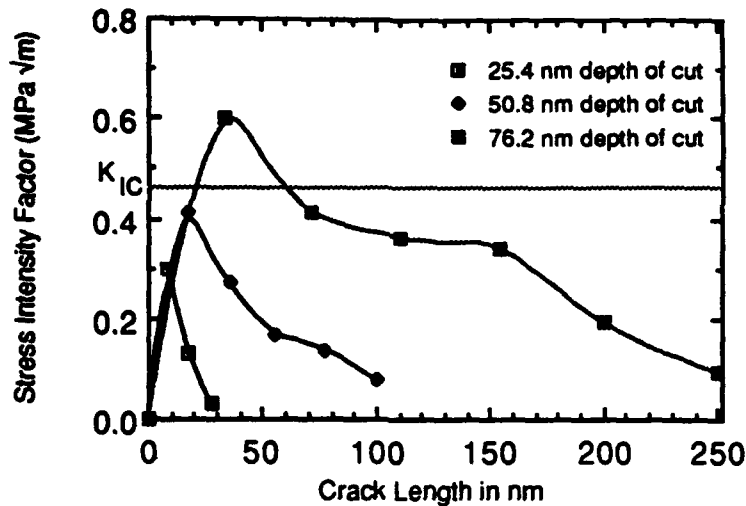
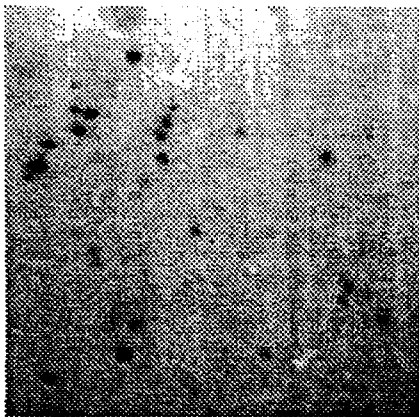


Figure 18: Variation of stress intensity factor with crack length for three depths of cut (germanium workpiece with a -10° rake angle).

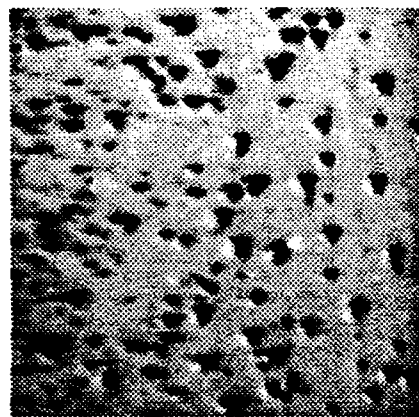
The area of the pits was calculated for each surface using the software program PrismView and compared to the total surface area to obtain a percent fractured area. Figure 20 shows the percent fractured area versus depth of cut. From this figure it appears that the critical depth of cut is between 38 and 50 nm, but closer to the 38 nm depth of cut. Thus, the simulation is consistent with these experimental findings.

9.4.2 Tool Forces

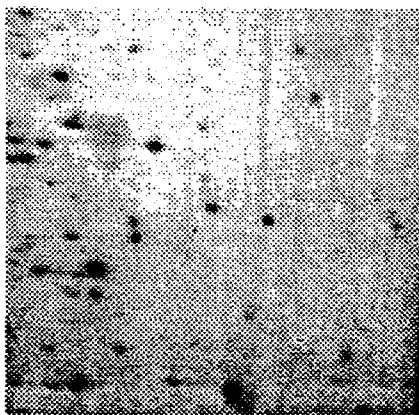
Tool forces in the cutting and thrust directions were measured for each depth of cut. Figure 21 shows the resulting curves. The thrust force is lower than the cutting force for each depth of cut, and the difference between the two forces increases with each increase in depth of cut. The forces sharply increase in going from the 25 to 38 nm depth of cut, but then decline for the larger depths of cut. It would appear that the curves should continue to increase for the larger depths of cut. It is suspected that because the larger depths show much more fracture, the forces are lower because the energy to fracture is much less than the energy required for plastic deformation. Therefore, these curves appear to verify that the critical depth of cut is between 38 and 50 nm, and closer to 38 nm.



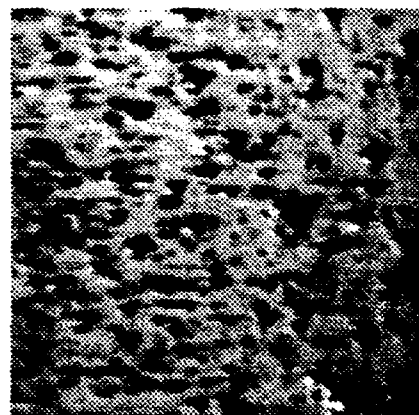
(a) 25 nm depth of cut (400 x 400 μm .scan).



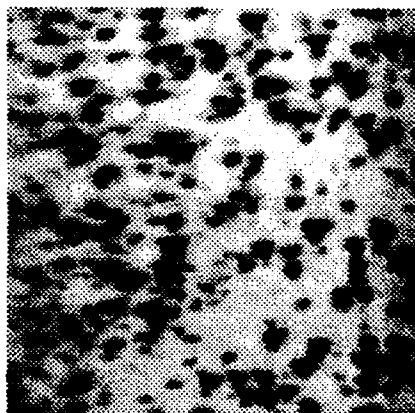
(c) 50 nm depth of cut (400 x 400 μm .scan).



(b) 38 nm depth of cut (400 x 400 μm .scan).



(d) 64 nm depth of cut (400 x 400 μm .scan).



(e) 76 nm depth of cut (400 x 400 μm .scan).

Figure 19: Machined surfaces for the 25, 38, 50, 64 and 76 nm depths of cut.

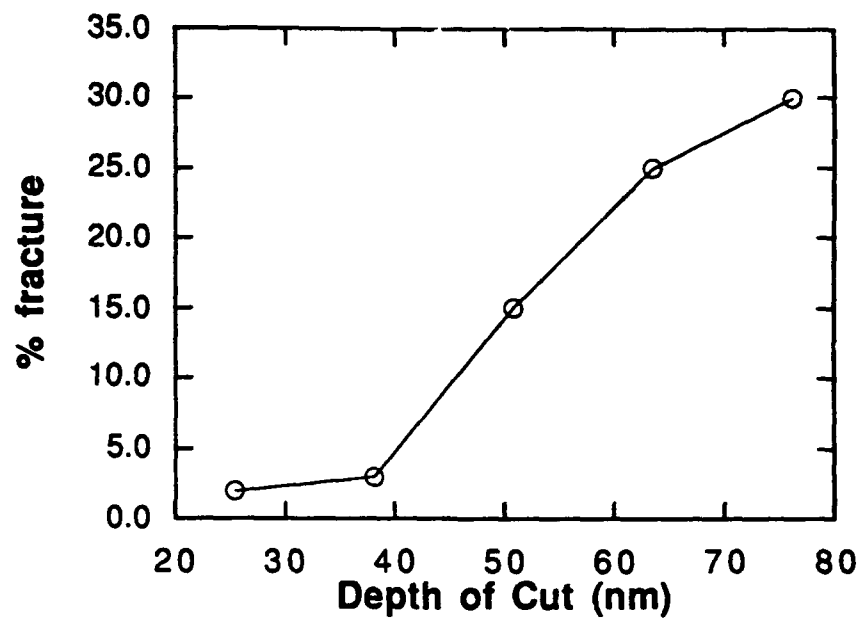


Figure 20: Graph showing percent fractured surface area for each depth of cut (germanium, 2.1 m/s, -10° rake angle).

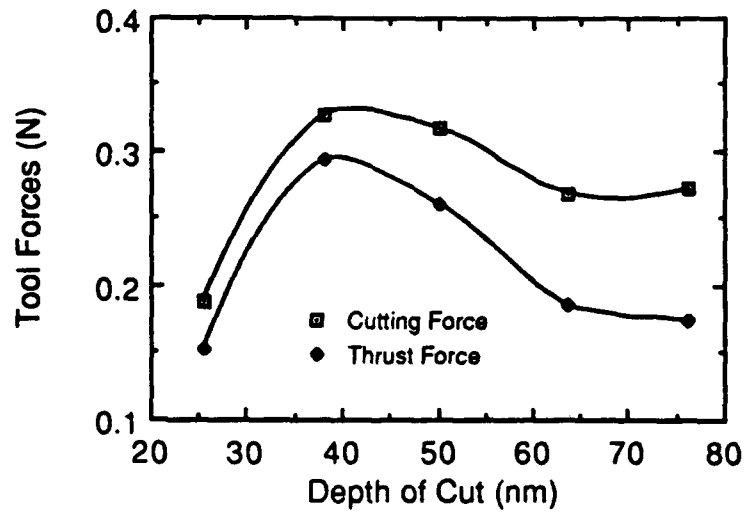


Figure 21: Experimental tool forces during the machining of germanium at various depths of cut

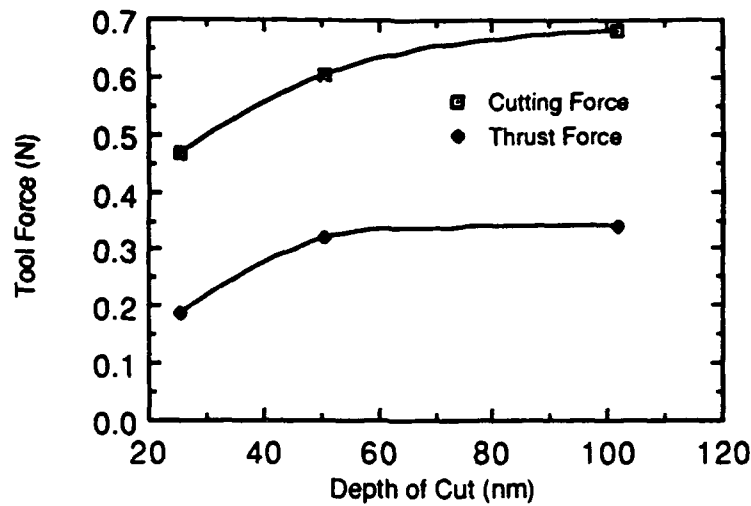


Figure 22: Predicted tool forces for the machining of germanium with a -10 degree rake angle.

Figure 22 shows the simulated tool forces for the -10 degree rake angle tool cutting germanium at depths of cut of 25, 50 and 100 nm. The simulated thrust force is less than the cutting force for each depth of cut which is similar to the experimental tool force relationship. However the simulated tool forces are greater than the experimental forces. Because fracture in each of the machined surfaces is ignored, the predicted tool forces are for completely ductile material removal. It has been shown [8] that the energy per unit volume of material removal is less for brittle fracture than for ductile machining. Therefore it is consistent that the forces predicted from the model would be larger than those measured if surface fracture was observed.

9.5 CONCLUSIONS

It was found that cutting speed had little effect on the surface damage from fracture. The highest speed simulated in the study, 16.8 m/s, did show a small reduction in the stress intensity factor suggesting that speeds higher than this could possibly affect fracture.

Work with simulating the cutting of silicon and germanium at different rake angles show that as the rake angle becomes more negative, the critical depth parameter becomes larger. This is because the tensile stresses do not extend as deeply into the surface, and as a result the stress intensity factor for a crack of a given length is reduced.

It was also found that increasing the tool edge radius lowers the stresses behind the tool, resulting in a similar effect to that of decreasing the rake angle.

A straight edge -10 degree rake angle, 6 degree clearance angle tool was chosen to machine germanium at 25, 38, 50, 64 and 76 nm depths of cut in order to more closely match the computer model and verify its predictions. The finite element model predicts a critical depth of cut of 50 nm. The measured surface fracture shows little pitting damage for the 25 and 38 nm depths of cut with fracture damage increasing for the 50 nm depth of cut and increasing greatly for the 64 and 76 nm depths of cut. This correlates fairly well with the model's prediction.

Tool forces were also measured for each depth of cut. The cutting force is larger than the thrust force for each depth of cut with the simulated tool forces showing the same relationship. The simulated tool forces are larger than the measured tool forces, but the simulated tool forces are for completely ductile material removal while the experimental tool forces were measured with fracture taking place in the larger depths of cut which would lower the tool force.

References

- [1] Hiatt, G., Strenkowski, J., "Fracture Mechanics Model for Predicting the Ductile Regime in Single Point Diamond Turning of Brittle Materials", *Precision Engineering Center 1990 Annual Report*, North Carolina State University, Raleigh, N.C., Vol. VIII, pp.181-200.
- [2] Asai, S. and Taguchi, Y. "Measurement of Cutting Edge of Diamond Tool", *Proceedings of Symposium at annual Meeting of JSPE*, 1988, pp.18-24.
- [3] Evans, C., Polvani, R., Postek, M., and Rhorer, R., "Some Observations on Tool Sharpness and Sub-Surface Damage in Single Point Diamond Turning", *SPIE*, Vol. 802, 1987, pp. 52-66.
- [4] Asai, S. and Kobayashi, A., "Observations of chip producing behavior in ultra-precision diamond machining and study on mirror-like surface generating mechanism", *Precision Engineering*, Vol. 12, No.3, 1990, p.139.
- [5] Asai, S., Taguchi, Y., Horio, K., Kasai, T., "Measuring the Very Small Cutting-Edge Radius for a Diamond Tool Using a New Kind of SEM Having Two Detectors", *Annals of the CIRP*, Vol. 39, No.1, 1990, p.88.
- [6] Miyamoto I., Kawata, K., Nishimura, K., and "Taguchi, Y., "Ion Beam Machining of Single Point Diamond Tools for Ultra-Precision Turning", *Journal of Materials Science Letters*, Vol. 7,,1988, p.1175.
- [7] Moriwaki, T., Okuda, K., "Machinability of Copper in Ultra-Precision Micro Diamond Cutting", *Annals of the CIRP*, Vol. 38, No. 1, 1989, p.118.
- [8] Bifano, T.G., and Fawcett, S.C., "Specific grinding energy as an in-process control variable for ductile-regime grinding", *Precision Engineering*, Vol. 13, No.4, 1991, pp.256-262.

10 DUCTILE-REGIME MACHINING OF GERMANIUM

R. M. Tidwell

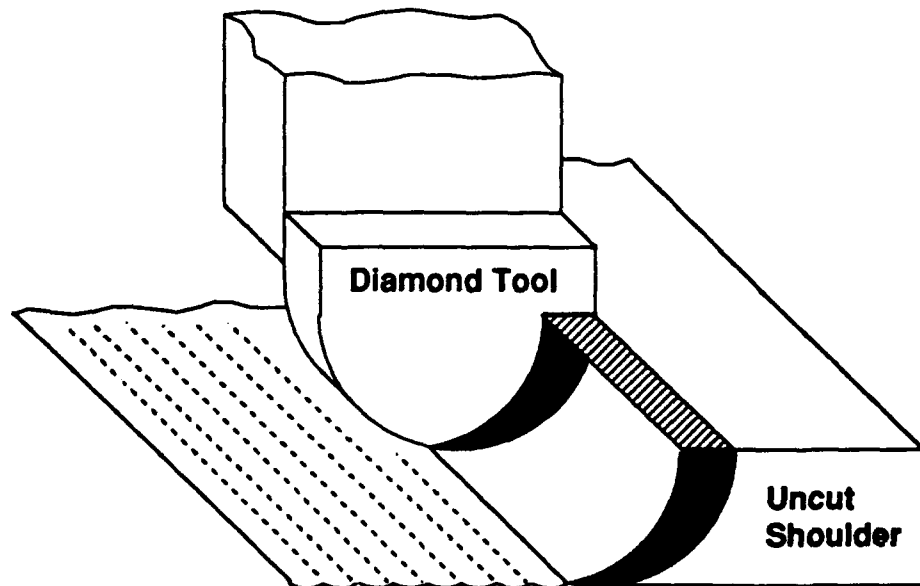
Graduate Student

R. O. Scattergood

Professor

Materials Science and Engineering

Interrupted cutting tests (ICT) were made on germanium single crystal wafers. Critical cutting depth d_c , mean crack size y_c , and limiting feedrates f_{max} for ductile-regime machining were obtained from the ICT data. A computer-based image analysis technique was used to process this data and provide better statistics than heretofore possible. The results showed that the improvements in machinability with increasing negative rake angles in germanium are due to an increase in d_c . The values of y_c were 5-10 times larger than d_c , and are not beneficial for machinability. Crystal orientation effects were also measured, and (100) surfaces showed the poorest machinability. A simplified stress model explains the rake angle effects, and also suggests that the large values of y_c are likely to result from subcritical crack growth in the unloaded state. Tool force measurements and surface finish measurements are also presented and discussed.



10.1 INTRODUCTION

One of the primary thrusts of the research on ductile-regime machining of germanium was the development of computer-based image analysis methods [1] to facilitate the interrupted cutting test (ICT) technique. The details of this test technique and the image analysis method have been discussed in an earlier report [2]. In essence, the ICT requires rapid withdrawal of the tool from the workpiece using a fast-tool servo. The position of the fracture damage transition point on the uncut shoulder is measured as a function of feedrate f . By suitable geometric analysis, the values of critical cutting depth d_c and mean fracture depth y_c shown in Figure 1 can be obtained from the data. The primary aim of the ICT is to measure d_c and y_c for given machining parameters. As shown in Figure 1, if the fracture damage of depth y_c , initiated at d_c , does not replicate below the cut plane of the finished surface, ductile-regime machining conditions have been achieved. This will occur even though substantial material removal still occurs by fracture processes.

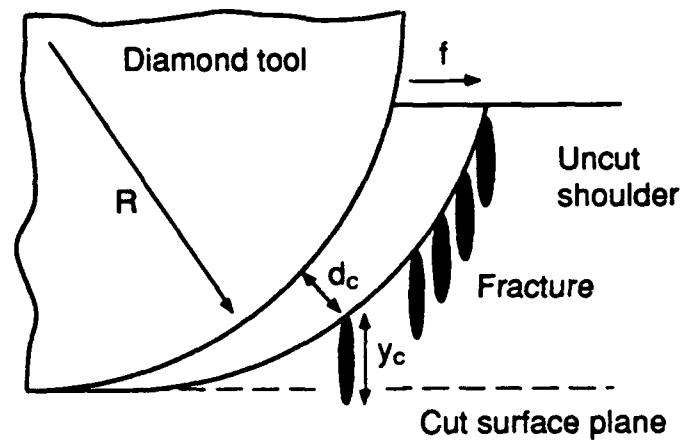


Figure 1: Schematic of single point diamond turning viewed along the cutting direction.

The feedrate f is a dominant machining parameter for achieving ductile-regime conditions. Small values of f move the ductile-to-brittle transition point at d_c higher on the tool nose. It is evident that there is a maximum allowable feedrate f_{max} at which fracture damage first replicates into the finished surface. This can be written as [1]:

$$f_{max}^2 = \frac{Rd_c^2}{d_c + 2y_c} \quad (1)$$

where R is the tool nose radius. d_c is the first-order parameter in this equation, with small values being undesirable, however, if y_c is large compared to d_c , it can also have a deleterious effect.

10.2 ICT EXPERIMENTAL RESULTS

ICT tests were made on single crystal germanium wafers having (100), (110) and (111) surface planes. The cutting directions in each surface plane were the directions of maximum fracture damage (worst-case conditions). Crystal geometry and details of the experimental procedures have been described elsewhere [1]. The primary machining parameters varied in the tests were the tool rake angle (negative 10°, 30° and 45°), and the coolant (dry - ambient air, and wet - distilled water). Tool nose radius $R = 3.175$ mm for all tests. The majority of the tests were conducted on (100) wafers. A limited number of test were made for tool-force measurements and surface finish measurements.

10.2.1 Rake Angle Effects

Figure 2 shows the values of d_c vs rake angle for dry and wet machining on (100) germanium wafer surfaces. There is a significant increasing trend in d_c with rake angle but coolant has only a minor effect. The general range of d_c for germanium is 50 - 300 nm for these test conditions.

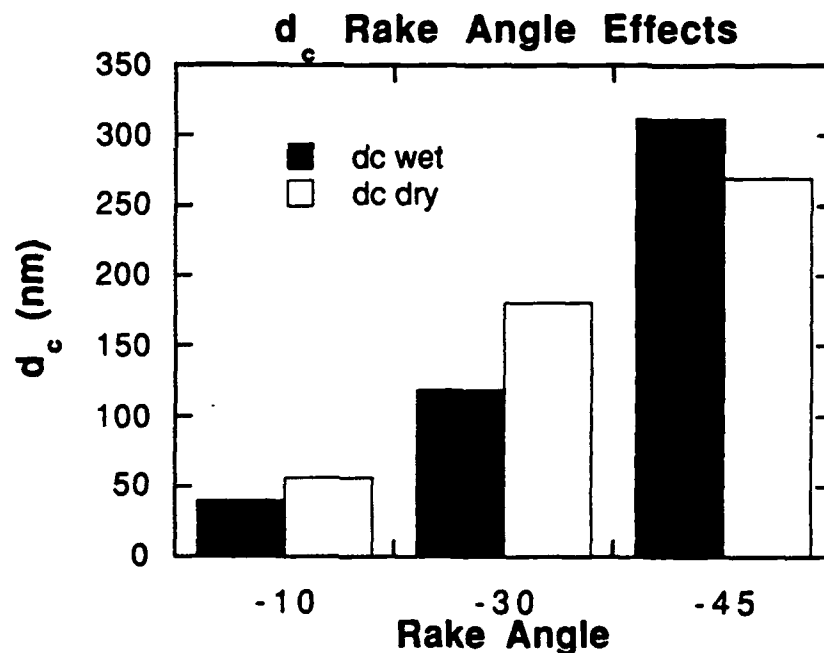


Figure 2: d_c vs rake angle for (100) germanium wafers.

Figure 3 shows the values of y_c vs rake angle for dry and wet machining on (100) germanium wafer surfaces. In contrast to Figure 2, there is no obvious trend in the y_c values, for either dry or wet machining conditions. Furthermore, compared to the d_c values, the values of y_c are 5 - 10 times larger.

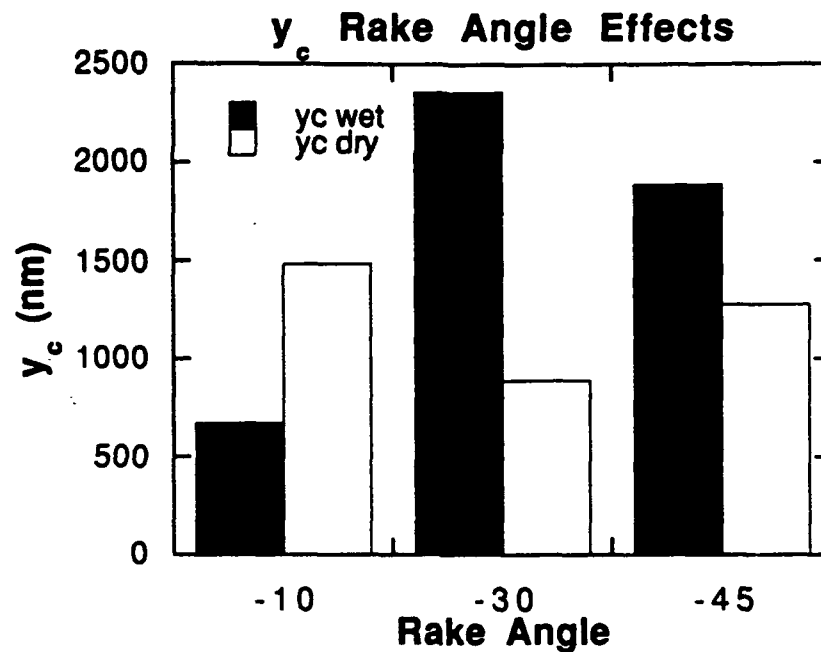


Figure 3: y_c vs rake angle for (100) germanium wafers.

Figure 4 shows the values of the limiting feedrate for ductile-regime conditions, f_{max} , obtained from Equation (1) vs rake angle for dry and wet machining on (100) germanium wafer surfaces. There is an increasing trend of f_{max} with increasing (negative) rake angle. In view of the results shown in Figures 2 and 3, it is clear that d_c is the first-order parameter in determining f_{max} , even though the large values of y_c cause a reduction in f_{max} as well as introduce some non-systematic scatter in the trend.

From the practitioners point of view, these results confirm the experience that increasing negative rake angle provides more "ductility" in the machining process for brittle materials. It is the large increase in d_c with increasing negative rake angle that produces the primary beneficial effect, manifest as the increase in f_{max} . In contrast, the large values of mean crack size y_c found in germanium do not favor ductile-regime conditions. In fact, if y_c were the order of d_c , Equation (1) shows that the f_{max} values in Figure 4 would be almost a factor of three larger.

10.2.2 Crystal Orientation Effects

The tests using different crystal surface orientations were done with a negative 30° rake angle tool. It must be emphasized that the tool cutting direction differed for each crystal surface orientation. The cutting direction was that direction in which surface fracture damage first occurred for each surface orientation (worst case condition). This was different for the (100), (110) and (111)

surfaces. The crystallographic nature of the orientation of surface fracture damage patterns have been described earlier [1]. Dry and wet machining conditions were studied for each of the different surface orientations.

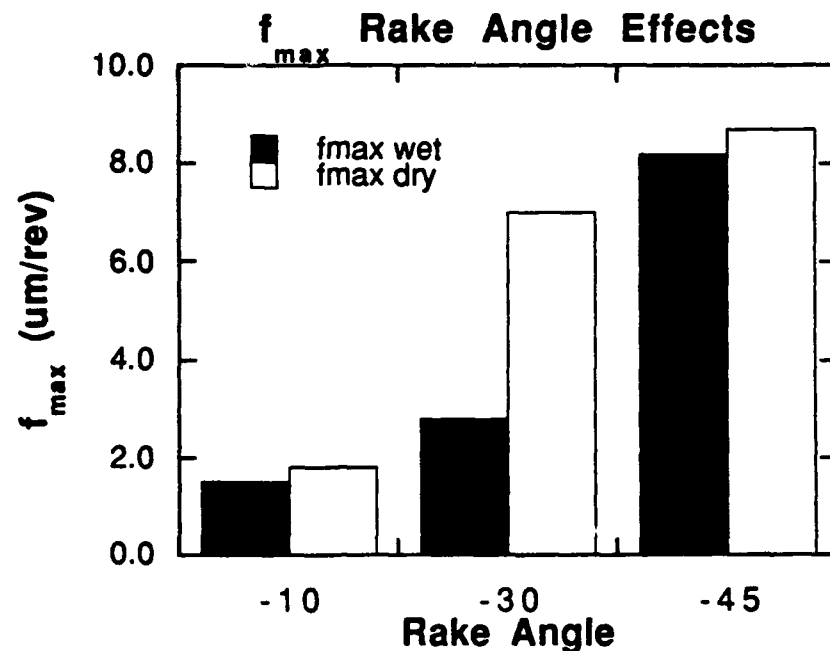


Figure 4: f_{\max} vs rake angle for (100) germanium wafers.

Figures 5-7 show d_c , y_c and f_{\max} as a function of the different crystal surface orientations for dry and wet machining conditions. The d_c values in Figure 5 are largest for (100) and (110) wafers; with dry conditions being somewhat better than wet. This same trend is apparent in the f_{\max} value in Figure 7, although the relatively large values of y_c for (100) wafers tend to negate somewhat the beneficial effect of d_c . Overall, the (110) wafers display the best ductile-regime machinability; that is, f_{\max} values are largest. In contrast, (100) wafers display the poorest machinability. This has implications for the machining of polycrystal germanium. The machinability in this case will be limited by grains with surface planes near the (100) orientation, especially under wet machining conditions. Dry machining should provide a noticeable improvement in performance for polycrystal material.

The ICT test results shown in Figures 2-7 are in good agreement with the results reported earlier by Blackley [3]. However, the ICT tests made by Blackley were analyzed using manual stereological methods rather than the computer-based image analysis methods used here, and so the scatter in the data was larger, especially for y_c values.

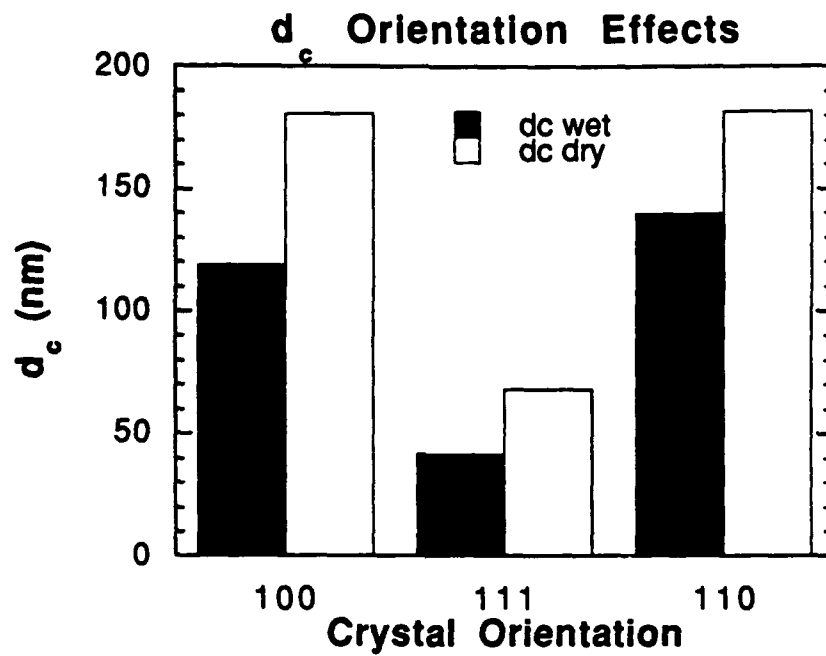


Figure 5: d_c vs crystal surface orientation for germanium wafers.

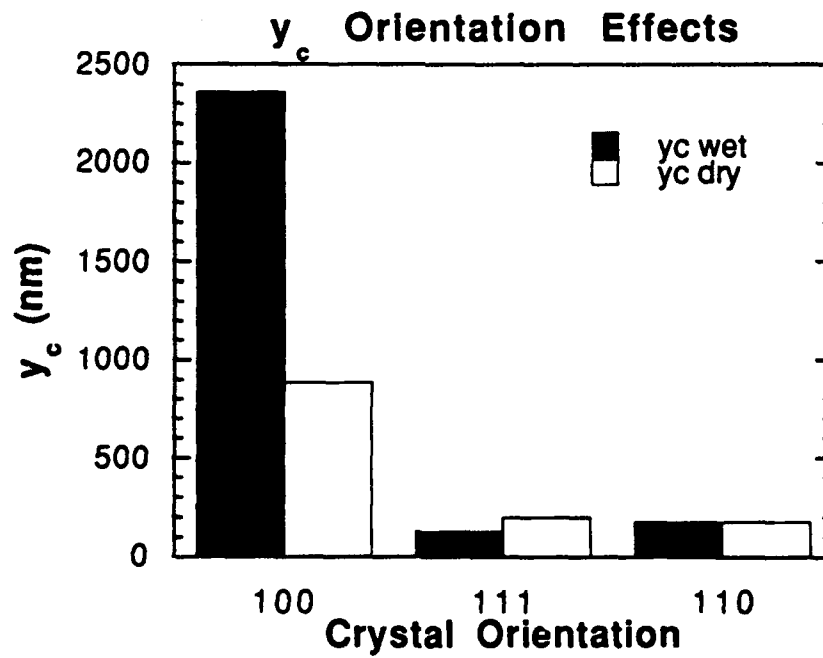


Figure 6: y_c vs crystal surface orientation for germanium wafers.

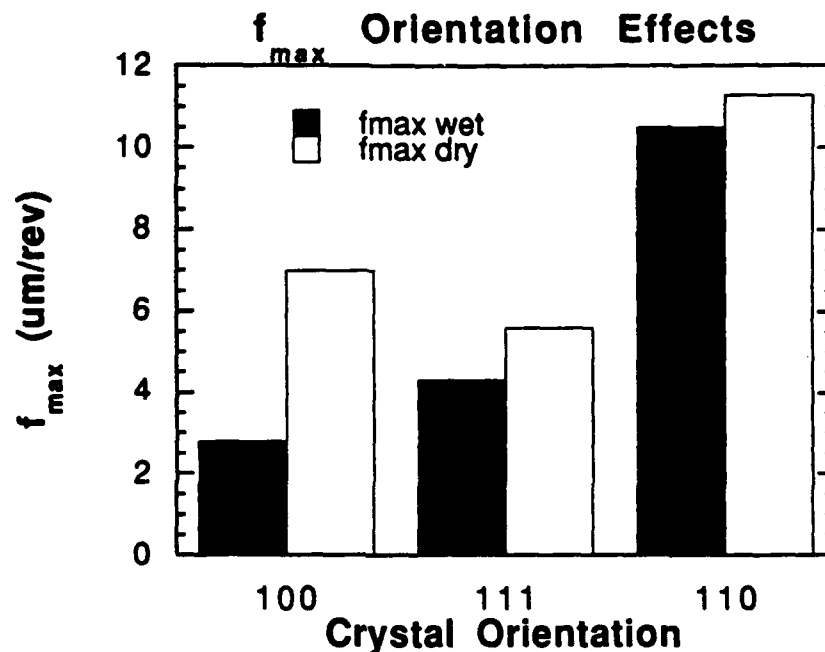


Figure 7: f_{max} vs crystal surface orientation for germanium wafers.

10.2.3 Tool Wear Experiments

Tool wear experiments were made using a negative 30° rake tool. A tool with several kilometers of testing on (100) germanium surfaces was compared to the same tool after 40 kilometers of testing. Rather surprisingly, the d_c and f_{max} values increased by about 40% after wear. The tool wear in this case was not severe, and appear to be limited to flank wear without serious degradation of the rake face or chipping of the cutting edge.

10.3 DISCUSSION OF ICT RESULTS

10.3.1 Elastic Stress Model

It appears that the beneficial effects of negative rake occur as a result of several factors. Rake angle effects on the critical cutting parameters d_c and y_c can be explained by a simplified model shown in Figure 8. This model depicts the stress contours due to a line force applied to an elastic half space. There is radial compression in front of the tool, i.e., in the region in front of the normal to the line force, and radial tension behind the tool. Plastic deformation will occur in shaded regions as shown, for example, using a Von-Mises yield criterion. In this model, fracture will occur in the tensile field, the crack trajectories being orthogonal to the stress contours because these are the

directions of maximum stress intensity for mode I (tensile) fracture. That this stress field is appropriate for precision machining has been confirmed in recent finite-element studies by Hiatt [4] (Section 9). Note that if cutting depths become very large, additional "wedging" forces may occur, similar to those used to analyze indentation fracture [5]. This will modify the stress field and the fracture morphology. Shear stresses could also lead to mode II fractures.

At any cutting depth, the stress intensity factor K has a maximum value K_{\max} , where the latter is the maximum value taken over all possible stress trajectories. Since K is bounded for localized stress fields, such a maximum will exist; that is, $K \propto \sigma\sqrt{c}$ and σ will decrease as $1/r$ where r is the distance from the force, while c will increase as \sqrt{r} so that $K \rightarrow 0$ in the limit as $r \rightarrow \infty$. For given force angle θ , P will increase proportionally with cutting depth d . Since the stress field scales with P , there will thus be a critical $d = d_c$ where $K_{\max} = K_c = \text{fracture toughness}$. Calculations based on this modeling approach have been made using simple analytical methods to evaluate the stress field. These will be described in detail elsewhere [6].

The new experimental data obtained here can be rationalized qualitatively from the picture presented in Figure 8. Increasing the rake angle to more negative values tilts the force P in Figure 8, i.e., increases the effective force angle θ . This in turn tilts the stress field, sweeping the tensile region up and out of the surface. A larger depth of cut is then possible since P must be increased to obtain the critical condition $K_{\max} = K_c$. The effect of environment on d_c (wet vs dry), while not large, is more difficult to rationalize. It could be a result of localized thermal heating, friction on the tool rake/flank, or any other effect that modifies P or θ . The changes are likely to be subtle in this context.

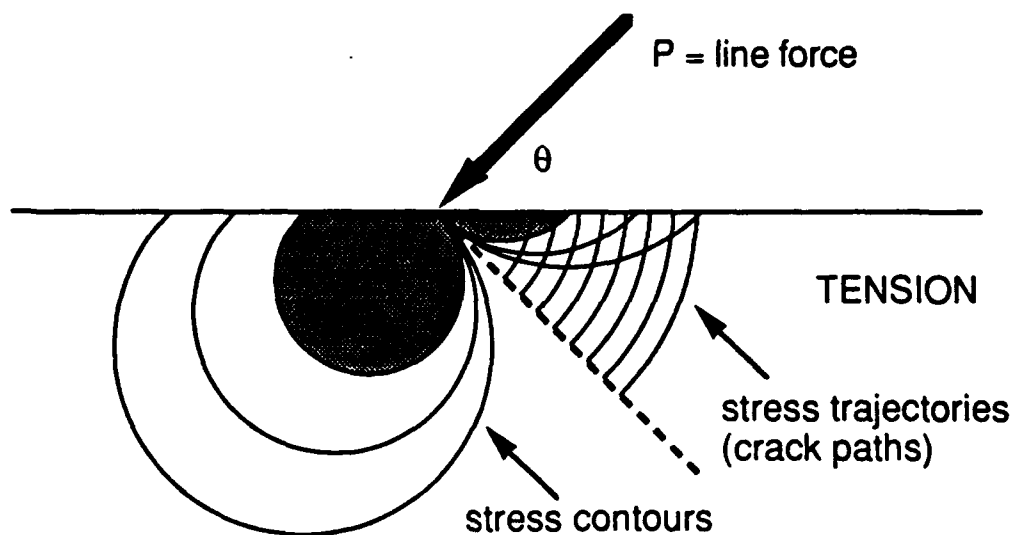


Figure 8: Schematic of the elastic-plastic stress field due to a line force P .

In the framework of the model shown in Figure 8, the large values of y_c (compared to d_c) and their often complex and non-systematic trends are unexpected. Finite-element calculations clearly show that d_c and y_c should be very comparable in magnitude, with y_c even being smaller than d_c [4]. That this does not occur suggests strongly that subcritical crack growth ($K < K_c$) occurs in the unloading field where the stresses change from the loaded state to the unloaded, residual stress state. Residual stresses present after machining are approximated by biaxial surface compression and biaxial subsurface tension. If cracks can "grow through" to the tensile field, substantial growth can occur leading to large values of y_c . Furthermore, wet environments may have a very pronounced effect on such subcritical growth. Slight changes in environment (water chemistry, humidity, etc.) will affect y_c in a subcritical growth mode, and this is one possibility for the the variability and large y_c values observed.

Perhaps surprisingly, a worn tool can improve machining performance. If wear causes significant rounding of the tool edge radius or a wear land on the flank face, the effective tool rake angle becomes more negative. This would be expected for cutting depths comparable to d_c , ie., for tool edge radius "rounding" the order of d_c . Since d_c is the order of 50- 300 nm, edge rounding on this scale is certainly possible. Finite-element calculations also reveal a potential beneficial effect of tool edge "rounding" [4]. The problem is complex, however, and diamond tool wear mechanisms are being evaluated by Larson (Section 8) and Drescher (Section 7).

Crystal orientation effects on damage have been noted earlier [7]. These will be difficult to analyze quantitatively for parameters such as d_c and y_c . Nevertheless, one should expect some variation. Earlier work [8] has shown that the critical depth d_c is given by

$$d_c = \beta \left(\frac{K_c}{H} \right)^2 \quad (2)$$

where β is a constant for fixed machining conditions. For brittle crystals like germanium, $K_c = \sqrt{2E\gamma}$ where γ = surface energy. Thus, there will be some orientation dependence of K_c due to anisotropy in surface energy. A complete analysis of the effect requires evaluating the $K_{\max} = K_c$ condition for d_c , taking into account the possible variation of K_c along differently oriented crack trajectories. In principle, this could be done using a Gibbs-Wulff construction for γ . The results are not self-evident, however, and the data reported here on orientation effects await a more detailed analysis.

10.4 TOOL FORCE AND SURFACE FINISH RESULTS

10.4.1 Tool Force Measurements

A limited number of tool force measurements were made using a negative 30° rake angle tool on (100) germanium wafers. The intent of these measurements was to explore possible correlations

between tool force and ductile-regime machinability. Tool force measurements provide one means for a real-time sensor to monitor the machining process. The results are shown in Table 1 for ambient air (dry) water (wet) and oil coolants. The ratio of thrust to cutting force is essentially the same in all cases, implying that the effective rake-face friction is independent of the coolant. However, the magnitude of the forces is noticeably different for each coolant; the cutting conditions and depth of cut were identical for each test. The force decreases for wet coolant compared to dry. This is in qualitative agreement with the results shown in Fig. 4 where f_{max} is seen to be significantly larger for dry cutting vs wet cutting at negative 30° rake. More fracture is expected for wet cutting, and this will lead to lower tool forces since fracture requires less energy than plastic flow for unit volume of material removed. On the other hand, oil coolant produces noticeably larger forces compared to either wet or dry conditions. This suggests that cutting with an oil coolant will enhance ductile-regime machining. Due to the difficulty in cleaning surfaces cut using oil, systematic ICT results were not obtained using oil. However, inspection of the uncut shoulders for selected ICT runs made with oil coolant showed much less fracture damage, relative to dry cutting, which is consistent with the expectations drawn from the tool-force measurements. Based on this limited set of data, tool force measurements, when appropriately calibrated, appear to be an attractive means for providing a real-time process monitor for ductile-regime machining.

	Dry	Wet	Oil
Cutting force (N)	.270	.236	.320
Thrust force (N)	.270	.238	.320

Table 1: Tool force results for (100) germanium.

10.4.2 Surface Finish Measurements

While not part of the major thrust of the research work, a preliminary study was initiated on surface finish measurements. The intent was to compare interferometer and AFM surface finish data on diamond-turned surfaces. Electroplated copper rather than germanium was used for these measurements, primarily to avoid difficulties encountered with oxidation of the germanium surfaces. The novel part of this work was the use of Fourier transform (FT) analysis to explore possible texture effects in the aperiodic, high frequency part of the surface finish. Because a single-point diamond cutting tool acts as a driven, non-linear system, one could expect some kind of unique fractal signature in the surface finish [9]. Furthermore, this would very likely differ in the cutting and transverse directions, thus producing an anisotropic surface texture.

Figure 9 shows an AFM scan of machined copper. The two-dimensional FT of typical AFM elevation data is shown in Figure 10. The sharp peaks in the FT represent the periodic part of the surface finish, i.e., tool feedrate markings. The diffuse background represents the high frequency aperiodic part.

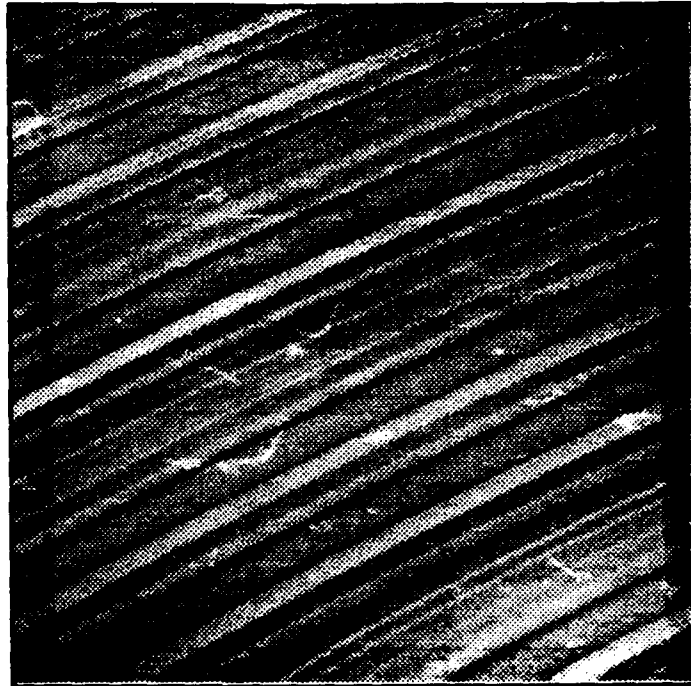


Figure 9: 60 μm x 60 μm AFM scan of machined copper (8 μm feedrate)

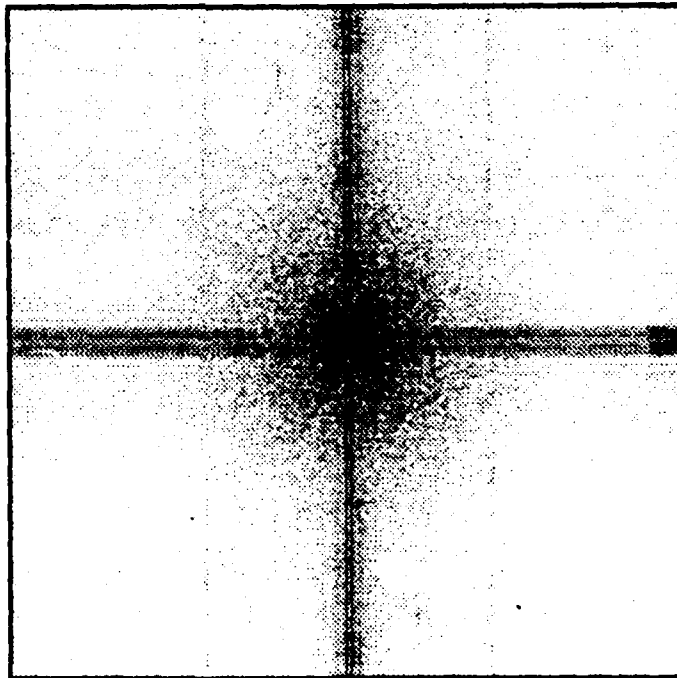


Figure 10: Two-dimensional FT of AFM elevation data for machined copper.

Using data like that shown in Figure 10, plots of log magnitude vs log frequency of the aperiodic part of the FT spectrum can be made for different directions in the machined surface. Linear plots were obtained and the fractal dimension of the linear surface trace can then be obtained from this slope [10]. By varying the direction of the linear surface trace, anisotropies in the fractal "texture" can be determined. Figure 11 is an orientation (Rose) plot of the fractal dimension obtained from AFM data on machined copper. The cutting direction is indicated by the arrow. There is a definite anisotropy in this data, with larger fractal dimensions being obtained in the cutting direction, as compared to transverse.

Data similar to Figures 9-11 were obtained on the same samples using a laser interferometer surface profiler. The most striking feature of these results was that the orientation plot was very similar to that shown in Figure 11. Furthermore, there was *quantitative* agreement for the values of the fractal dimension. This suggests that the aperiodic texture in the machined surfaces retains its fractal characteristics over a large range of measurement scale (the spatial in-plane resolution differs by about three order of magnitude for AFM vs interferometer data). While these results are preliminary, they strongly suggest that the machining process replicates a definite, anisotropic fractal texture in the surface. Further work is needed to characterize this texture and correlate it with machining parameters. This kind of analysis could provide useful information on machine dynamics and material removal mechanisms, both for ductile and brittle materials.

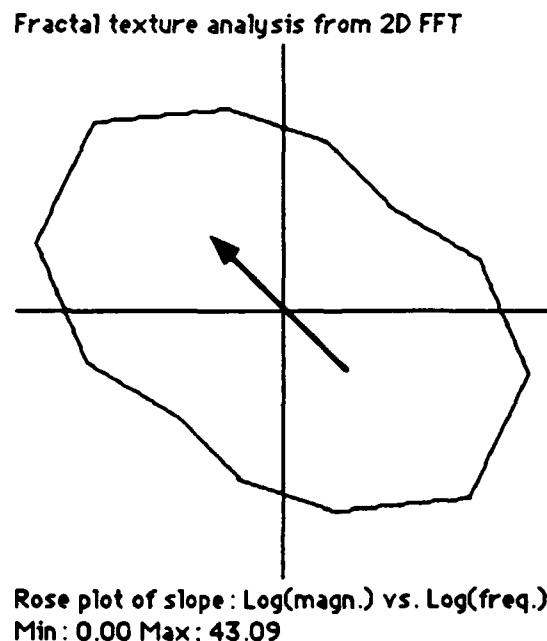


Figure 11: Orientation (Rose) plot of the fractal dimension obtained from AFM data.

References

1. R. M. Tidwell, MS Thesis, North Carolina State University (1992).
2. Precision Engineering Center Annual Report, Vol. VIII, p. 153 (1990).
3. W. S. Blackley and R. O. Scattergood, *Ductile-Regime Machining Model for Diamond Turning of Brittle Materials*, Prec. Engr., Vol. 13, No. 2, 97 (1991).
4. G. Hiatt, PhD Thesis (in progress), North Carolina State University (1992).
5. R. F. Cook and G. M. Pharr, *Direct Observations and Analysis of Indentation Cracking in Glasses and Ceramics*, J. Am. Ceram. Soc., Vol. 73, No. 4, 787 (1990).
6. R. O. Scattergood, R. M. Tidwell and W. S. Blackley, to be submitted for publication (1992).
7. W. S. Blackley and R. O. Scattergood, *Crystal Orientation Dependence of Machining Damage*, J. Am. Ceram. Soc., Vol. 73, No. 10, 3113 (1990).
8. T. G. Bifano, PhD Thesis, North Carolina State University (1988),
9. P. Scott, *Some Recent Developments in the Analysis and Interpretation of Surface Topography*, Proc. ASPE 1990 Ann. Conference, 84 (1990).
10. J. Feder, *Fractals*, Plenum Press, New York (1989).

11 CONTOUR GRINDING OF BRITTLE MATERIALS

G. Walter Rosenberger

Graduate Student

G. McDonald Moorefield II

Research Assistant/Lecturer

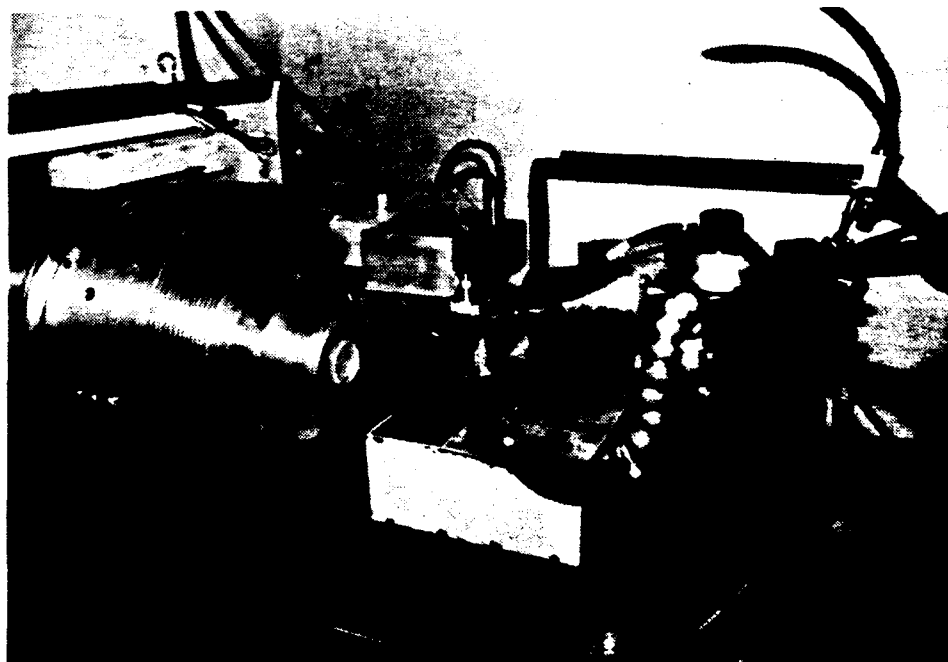
Dr. David A. Dickey

Professor, Department of Statistics/Economics and Business

Dr. Thomas A. Dow

Professor, Department of Mechanical and Aerospace Engineering

A statistically designed set of experiments was undertaken to evaluate the geometrical and operational parameters associated with the grinding of brittle materials in the ductile mode. Sixteen specimens were ground under controlled conditions wherein 12 parameters were specified at two distinct values. The specimens were examined, measured for surface roughness, and compared to analytical predictions for surface finish based on geometric modelling. The initial set of experiments identified several parameters which merit further study. Wheel characteristics, coolant chemistry and material properties, and geometric parameters will receive primary focus.



11.1 INTRODUCTION

One of the major limitations of grinding on brittle materials such as ceramics and glass is the subsurface damage introduced by the grinding operation. Fracture damage is particularly undesirable in virtually all finished components. Significant post-polishing is typically required to remove the damage. Since polishing is costly and can produce figure errors for some surfaces, there is a definite advantage in being able to eliminate or control the fracture damage introduced by grinding. Total elimination of the damage in a finished surface is termed "ductile regime grinding." This represents optimum performance for grinding as a fabrication technique.

The work described in this report attempts to develop a more fundamental understanding of the operational parameters that control ductile-regime grinding processes. Current production methods rely on trial and error for the selection of conditions which are thought to be optimal. A competitive technology must go beyond this iterative approach; therefore, an improved level of understanding of basic phenomena is necessary.

Prior research work at the PEC has led to a new understanding of material response in machining and grinding operations. The primary consideration in the successful implementation of any ductile-regime process is the location and ultimate control of the transition point d_c on the uncut chip profile.

The critical cutting depth d_c can be expressed as [1]

$$d_c = \beta \left[\frac{K_c}{H} \right]^2 \quad (1)$$

where the critical material factor is the ratio of the fracture toughness K_c to the hardness (plastic flow resistance) H . The parameter β is determined by operational parameters which include wheel bond, grit size, coolant, etc. The specific dependence of β on these parameters is not currently well enough understood to be derived from first principles, but rather must be obtained from appropriately designed experiments.

11.2 CONTOUR GRINDING EXPERIMENTS

A set of statistically designed experiments was undertaken following the sub-factorial methodologies of Plackett-Burman. The objective of these experiments was to determine the existence of functional relationships between surface finish and various parameters associated with the grinding of brittle materials. The statistical design approach has a significant advantage in that a maximum amount of information can be gained from a relatively small number of experiments.

The method is highly flexible and can be implemented to check interactive effects, non-linear effects and experimental reproducibility.

Twelve parameters were examined, including grinding wheel characteristics, workpiece material properties, coolant, and geometrical parameters such as depth of cut, feed rate and rotational speeds. The parameters for each of the 16 experiments are listed in Table 1.

Test No.	Cool	N_w	Mat'l	Conc	N_p	D_p	A	Diamond Size	Dress Tech	d	Feed	Bond
1	water	50	Zer	50	100	12	1	30-60	nib	6	0.5	Metal
2	water	30	Zer	50	25	6	1	4-6	nib	3	0.15	Resin
3	water	30	SiC	100	25	12	1	30-60	stick	6	0.5	Resin
4	water	50	SiC	100	100	6	1	4-6	stick	3	0.15	Metal
5	oil	30	SiC	100	100	12	1	4-6	nib	6	0.15	Metal
6	water	30	SiC	50	25	12	5	4-6	stick	3	0.5	Metal
7	oil	50	SiC	50	25	6	5	4-6	nib	6	0.5	Metal
8	oil	50	SiC	100	25	6	1	30-60	nib	3	0.5	Resin
9	oil	30	Zer	100	100	6	5	4-6	stick	6	0.5	Resin
10	oil	50	Zer	100	25	12	5	30-60	stick	3	0.15	Metal
11	water	50	Zer	100	100	12	5	4-6	nib	3	0.5	Resin
12	oil	30	SiC	50	100	12	5	30-60	nib	3	0.15	Resin
13	water	50	SiC	50	100	6	5	30-60	stick	6	0.15	Resin
14	oil	50	Zer	50	25	12	1	4-6	stick	6	0.15	Resin
15	water	30	Zer	100	25	6	5	30-60	nib	6	0.15	Metal
16	oil	30	Zer	50	100	6	1	30-60	stick	3	0.5	Metal

(Cool: Coolant, either water or oil; N_w : Wheel Speed (rpm x 1000); Mat'l: Part Material, either Zerodur or Silicon Carbide; Conc: Wheel Diamond Concentration; N_p : Part Speed (rpm); D_p : Part Diameter (mm); A: Wheel Nose Radius (mm); Diamond Size: Wheel Diamond Size (μ m); Dress Tech: Wheel Dressing Technique, either programmed diamond Nib truing or manual Stick dressing; d: Depth of Cut (μ m); Feed: Cross feed rate (mm/min); Bond: Wheel Binder Material, either Resin or Metal.)

Table 1: Grinding conditions for the sixteen experiments.

Two states, a *high* and a *low*, were set for each parameter. The 16 experiments yielded a set of ground surfaces of which 8 had each variable at a high value and 8 had each variable at a low value. The effects of the *high/low* parametric changes were evaluated by measuring the Arithmetic or Average Surface Roughness (R_a) for 16 samples ground under controlled experimental conditions. A regression analysis of the results allowed a first-order approximation of the significance of that change in variable.

11.2.1 Statistical Design of Experiment

Statistical analysis was used to identify correlations between the measured R_a values and the two states for each of the twelve parameters. The two states selected for each parameter were presumably different enough to characterize the effect of changing that particular parameter. Arbitrarily one condition would be assigned "1" (=high) and the other condition would be assigned "-1" (=low). Table 2 is the experimental matrix showing the parameter states.

Test No.	A	B	C	D	E	F	G	H	I	J	K	L
1	-1	1	-1	-1	1	1	-1	1	-1	1	1	1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	1	1	-1	1	-1	1	1	1	1	-1
4	-1	1	1	1	1	-1	-1	-1	1	-1	-1	1
5	1	-1	1	1	1	1	-1	-1	-1	1	-1	-1
6	-1	-1	1	-1	-1	1	1	-1	1	-1	1	1
7	1	1	1	-1	-1	-1	1	-1	-1	1	1	-1
8	1	1	1	1	-1	-1	-1	1	-1	-1	1	1
9	1	-1	-1	1	1	-1	1	-1	1	1	1	1
10	1	1	-1	1	-1	1	1	1	1	-1	-1	-1
11	-1	1	-1	1	1	1	1	-1	-1	-1	1	-1
12	1	-1	1	-1	1	1	1	1	-1	-1	-1	1
13	-1	1	1	-1	1	-1	1	1	1	1	-1	-1
14	1	1	-1	-1	-1	1	-1	-1	1	1	-1	1
15	-1	-1	-1	1	-1	-1	1	1	-1	1	-1	1
16	1	-1	-1	-1	1	-1	-1	1	1	-1	1	-1

Table 2: Orthogonal array of *high* and *low* states.

A single representative equation can be written to describe the effects of the various parametric changes on the resulting surface finish:

$$R_a = K + C_1*A + C_2*B + C_3*C + \dots + C_{12}*L \quad (2)$$

where K is a constant, C_1 through C_{12} are the parametric coefficients, and A through L are the parameter states (either "1" or "-1"). The constant and the coefficients can be determined by performing a least-squares regression, such that the above equation defines the surface roughness in terms of the respective parameters for a particular experiment. That is, by inserting the parameter states, the equation will yield the surface roughness measured for a particular sample.

11.2.2 Grinding Procedure

Equipment The sixteen grinding experiments were performed on the PEC DTM (Rank Pneumo ASG2500). Figure 1 shows the grinding setup. The relative motion necessary for machining is provided by two hydrostatic bearing slides oriented in a "T-base" configuration, with the air bearing workpiece spindle mounted to the Z-axis slide and the ancillary grinding spindle mounted to the X-axis slide.

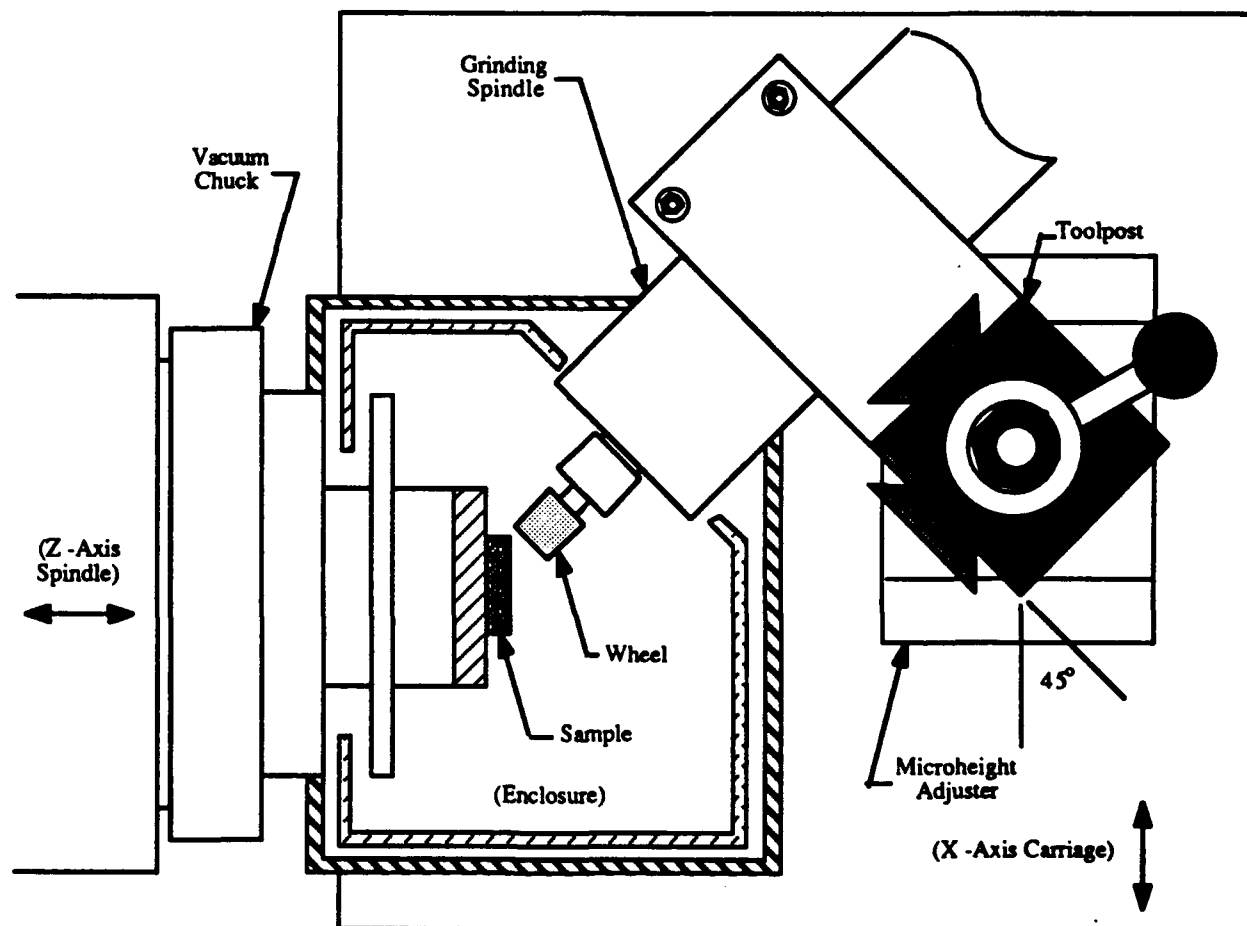


Figure 1: Equipment setup for contour grinding experiments.

A Rank Pneumo Microheight Adjuster was mounted to the X-axis of the ASG2500; to the top of this unit an Aloris Dovetail Toolpost was attached. The grinding spindle was clamped in a dovetail mounting bracket, then locked onto the Aloris Dovetail Toolpost so that the spindle axis was horizontal, oriented at a 45° angle to the X-axis slide (see Figure 1). The grinding spindle was a Federal Mogul Westwind D1090 air bearing, air turbine, low torque, light load spindle capable of speeds up to 60,000 rpm. The spindle was operated open loop, but the speed was monitored using an OSSEM Optical Detector¹ to detect twenty inked graduations on the spindle rotor. A Fluke 1900A Multi-Counter was used to count the graduations detected by the optical detector. An enclosure (see Figure 1) was constructed to contain the coolant spray from the rotating grinding wheel. In addition, a special fixture was fabricated to vacuum mount the test specimens on the spindle. This mount provided repeatable positioning of the test samples. The center hole in the sample mounting fixture also held the diamond dressing nib used for the wheel truing/dressing procedure discussed later.

Each of the eight grinding wheels was cylindrical, 12.5 mm in diameter and 12.5 mm long, with an integral 6.25 mm diameter carbide mandrel. The wheel composition/preparation variables are listed in Table 3. The wheels were purchased from Diagrind, Inc.

Wheel Composition Variables:			Wheel Preparation Variables:	
Binder Material:	Diamond Concentration:	Diamond Size: (Grit)	Wheel Dressing Technique:	Wheel Nose Radius:
Metal	50	4-6 μm (3000)	Programmed Nib Truing	1 mm
or	or	or	or	or
Resin	100	30-60 μm (400)	Manual Stick Dressing	5 mm

Table 3: Wheel Composition/Preparation Variables.

Two workpiece materials were studied: zerodur and chemical-vapor-deposited (CVD) silicon carbide. These materials were selected because their critical chip thickness values differ by nearly an order of magnitude. (The critical chip thickness for Zerodur is approximately 70 nm as compared to approximately 600 nm for silicon carbide [2].) The 25.6 mm diameter, 3.3 mm thick Zerodur samples were purchased from Schott Glass Technologies, Inc. The 12 mm diameter, 8 mm thick silicon carbide samples were provided by the Kodak Apparatus Division of Eastman Kodak Company. Each sample was cemented to a serialized aluminum base using cyanoacrylate.

¹ Manufactured by Opto Acoustic Sensors, Inc.

Procedure The specified wheel nose radius was dressed on the front wheel edge (see Figure 2). The stick dressing technique used an aluminum oxide stick for dressing resin bond wheels and a silicon carbide stick for dressing metal bond wheels. Stick dressing involved manually shaping the desired radius on the edge of the wheel, rotating at approximately 60,000 rpm. A Questar Long Range Microscope equipped with a Panasonic Video Camera and Monitor was used to view the dressing process. Two scaled, transparent templates were made to provide an inspection of the wheel radius as viewed on the monitor at a magnification of 43.8 times the actual wheel size.

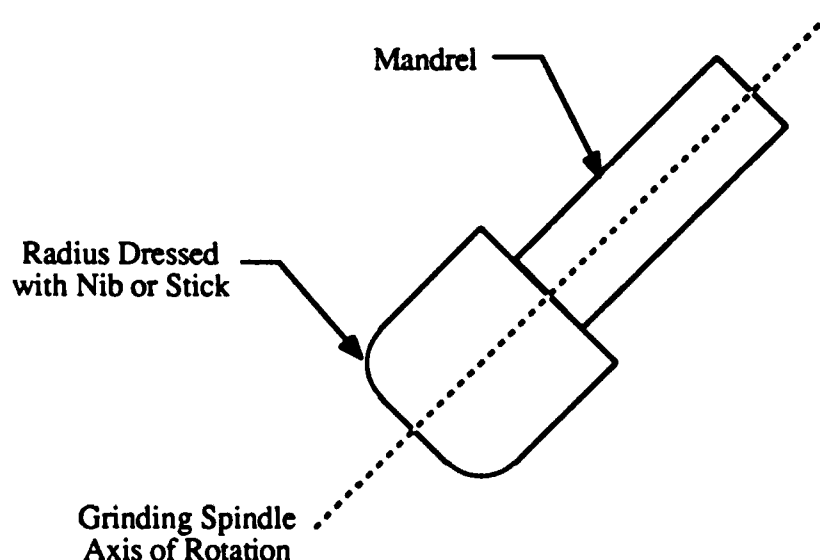


Figure 2: Grinding Wheel Geometry.

Nib truing involved machining the desired radius on the edge of the rotating wheel using an industrial diamond nib. The fixture design required that the sample be removed from the mounting fixture and the 9.5 mm diameter shank of the truing nib be secured in the fixture's central hole. The workpiece spindle was locked to prevent rotation of the nib. The lathe was programmed to generate the desired nose radius on the wheel. Nib truing was performed with the spindle rotating at approximately 60,000 rpm, using the coolant dictated by the respective test conditions.

Two types of coolant were used, water and oil. The water coolant was ordinary tap water. The oil coolant was Mobilmet Omicron, a refined mineral oil cutting fluid with a viscosity of 26.6 centistokes at 38°C. Each coolant was applied to the vicinity of the wheel/sample interface in a constant stream, at approximately room temperature (20°C). The coolants were not recirculated. The flow rate of the coolant was not measured.

The sample's surface was "rough" ground prior to the experimental cut using the same values for the parameters (wheel, wheel and part speeds, and coolant) as called out in the experiment. The rough grinding pass used a 5 μm depth of cut and a crossfeed rate of 2.5 mm/min. Next, the wheel was trued or dressed according to the experimental specifications. After completion of the wheel preparation, the sample was replaced on the mounting fixture (if nib truing was used), the depth of cut, wheel and part speeds were set, the necessary movements were programmed, the coolant stream was directed toward the grinding region, and the surface was ground. The experimental grinding cut began at the specified part diameter and proceeded 3 mm towards the center of the part. Conventional grinding, where the wheel and part tangential velocities are opposite, was used for all the tests.

The surfaces were cleaned using several cleaning agents, including acetone, ethanol, methanol, and Freon. The test surfaces were flooded with cleaning fluid, then drag-cleaned with tissues.

11.2.3 Measurement Procedure

Each test surface was examined with three instruments: a Zeiss Nomarski Microscope, a Talysurf Stylus Profilometer, and a Zygo Maxim 3D Laser Interferometric Microscope.

Zeiss Nomarski Microscope The test surface was first inspected with the Zeiss Nomarski Microscope. During this inspection, the nature of the surface finish (whether the surface was characterized by grinding feeds or by fractures and pits) was determined. Most often the surface was inspected at 500X magnification.

Talysurf Stylus Profilometer The actual depths of cut were measured using a Talysurf 5 Profilometer System. The 6 μm radius diamond stylus of the Talysurf was traversed radially across the test region (from one shoulder to the other shoulder). The depth of cut was interpreted as the vertical separation of the roughed surface and the test surface, as measured at the outer shoulder. The Talysurf was also used to measure the Arithmetic Average Roughness (R_a) of the ground test region. The R_a measurement involved performing five radial line scans at the outer shoulder of the test region.

Zygo Maxim 3D Laser Interferometric Microscope The Maxim 3D system is built around a Helium-Neon Laser and can be fitted with Fizeau and Mirau microscope objectives of various magnifications. All measurements were recorded using the 100X Fizeau Objective. Five areas were selected within the test region which represented the sample's best quality surfaces. An R_a value was calculated for each 78 μm X 57 μm view area. Surfaces characterized by pits and fractures produced phase discontinuities, resulting in regions which were unresolvable by the Maxim 3D. Such regions were removed by the Zygo software and the subsequent R_a calculations.

This calculation procedure may be part of the reason for the differences in surface roughness between the two profilometers.

Figures 3 and 4 show the output from the Maxim microscope for two ground samples. The dominant features from Sample 8 in Figure 3 appear to be feed markings (spaced at $37.6\text{ }\mu\text{m}$) and the dark pitted areas. The pits appear to be fracture sites where the interference fringes are too closely spaced for the microscope to resolve. In Figure 4, Sample 11 shows low surface roughness and minimal surface features.

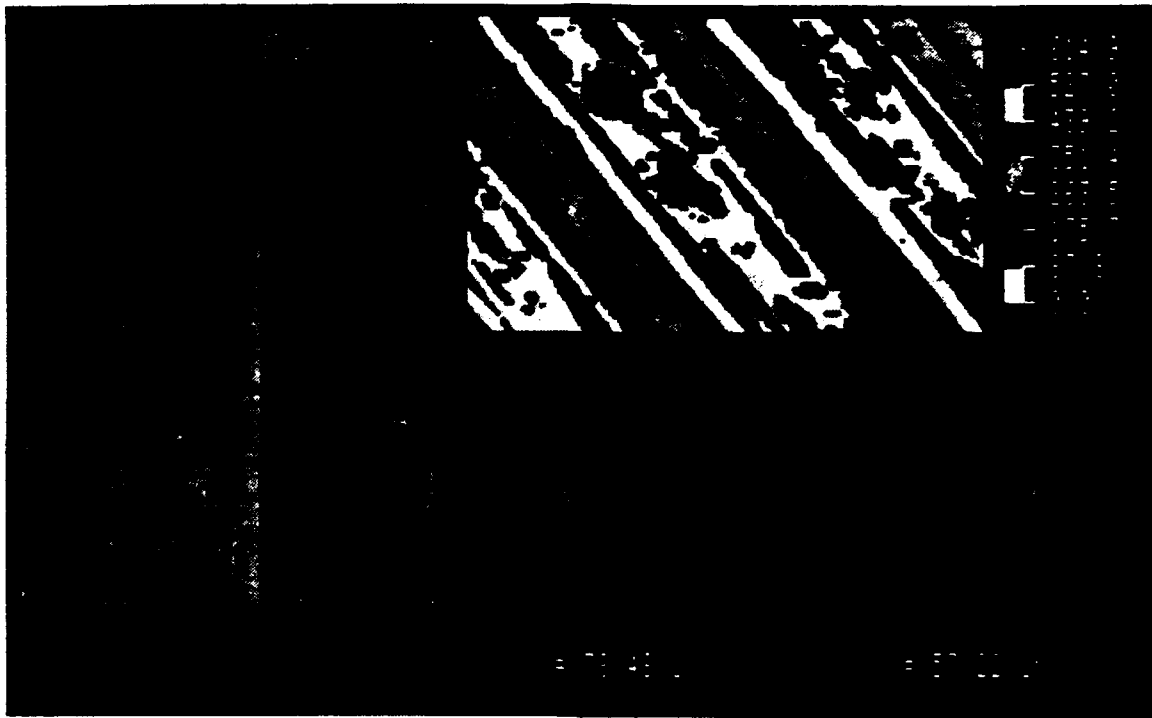


Figure 3: Surface of Sample 8 examined by the Maxim 3D Microscope

11.2.4 Data Analysis Procedure

Each set of five R_a values was processed to obtain a single representative value for that set. Several procedures were used: averages were calculated, averages were calculated with the high and low values removed, ranges were calculated, and maximums were identified. The various representative values were entered as a column in the orthogonal state array. Two arrays were constructed, one for the Talysurf data and one for the Zygo data. A SAS Institute statistical analysis package was used to perform the linear regression on the array to determine the parametric coefficients in the representative equation (Equation (2)). The output of the analysis was examined to determine the presence of any significant trends.

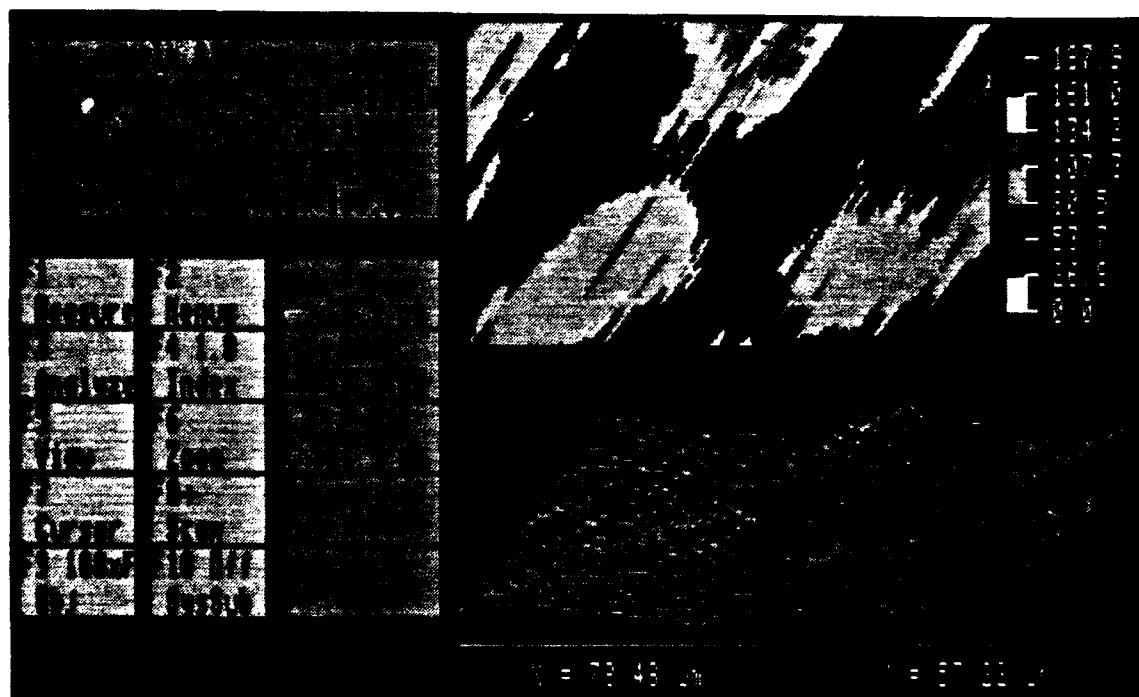


Figure 4: Surface of Sample 11 examined by the Maxim 3D Microscope

11.2.5 Results and Discussion

The results of the experiments are summarized in Figure 5. This figure shows the value of the coefficients calculated for each variable in the sequence of experiments in the form of Equation (2). Positive values mean that the surface roughness increased when the variable was changed from its low state (-1) to its high state (+1) and negative values indicate a reduction in surface roughness as a result of changing from the low to high state. The magnitude of the change in R_a is twice the value of the coefficient.

The results from this series of experiments require a disciplined interpretation. Because all of the parametric changes have been assigned equal weight (" +1" or " -1"), regardless of the magnitudes of the changes, general comparisons of the coefficients can be misleading. For example, the low to high changes represent a 900% increase in diamond grit diameter, but only a 100% increase in diamond concentration or part diameter. Variables such as coolant and dressing technique are discrete and a magnitude comparison is not possible. Furthermore, known interactions between parameters were not addressed. The limited number of experiments and large number of parameters filled the experimental array and did not leave room in the parameter space to gain statistical information about interactions. The Coefficient of Determination (COD) for the various

sets of values ranged from 0.73 to 0.82, indicating reasonable correlation between the predicted and actual R_a values. Since the results for the four types of representative R_a values showed excellent agreement, only the results obtained using the average R_a values are shown.

Two factors could explain why the Talysurf readings are consistently larger than the Zygo readings. First, the Talysurf values were based on a linescan while the Zygo values were calculated for a sample area. The second reason is that the R_a values calculated by the Zygo did not include the effects of pits, fractures, etc., because of phase discontinuities.

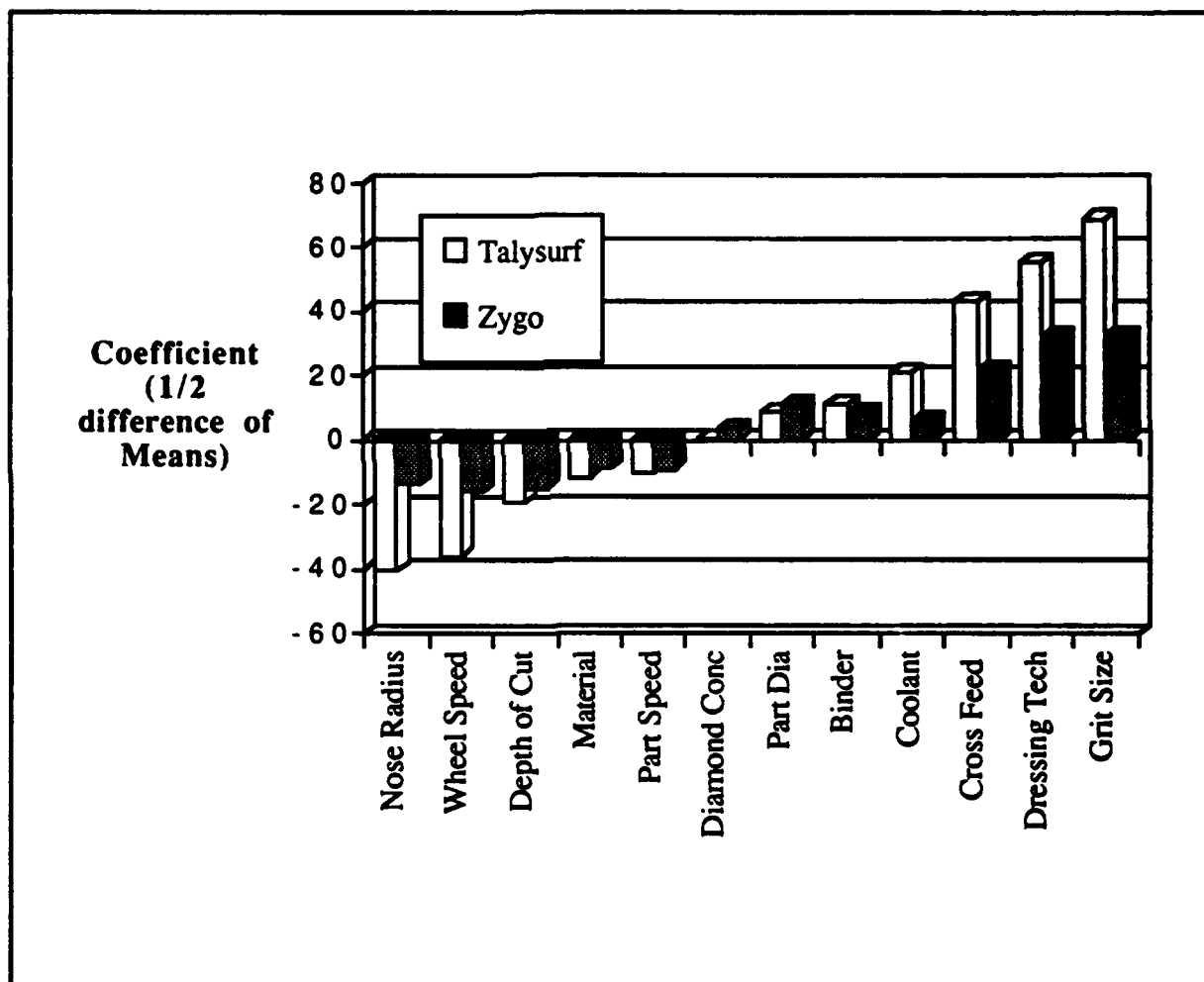


Figure 5: Experimental parametric coefficients (change in R_a value, in nm).

The two most influential parameters, in terms of surface finish, were the changes in the diamond particle size and the dressing technique. Changing the size of the diamonds had the largest influence on surface finish (The Zygo data coefficient was 68.9 and the Talysurf data coefficient

was 32.4.), with the smaller diamonds producing lower values of R_a and the larger diamonds producing higher R_a values. Changes in the wheel dressing method produced the next highest level of influence (The Zygo data coefficient was 55.4 and the Talysurf data coefficient was 32.0.), with nib dressing resulting in lower R_a values and stick dressing resulting in higher R_a values. These findings could support a hypothesis that the dressing technique has a significant impact on the diamond particle size as well as the condition of the diamonds and the geometry of the wheel.

Other significant coefficients were found for the crossfeed rate (Zygo: 21.9; Talysurf: 43.0), wheel nose radius (Zygo: -13.5; Talysurf: -40.4), and wheel speed (Zygo: -15.7; Talysurf: -36.0). As expected from the geometric model for theoretical surface finish [4], the larger crossfeed rate, smaller wheel nose radius, and lower wheel speed all produced higher values of R_a while the smaller crossfeed rate, larger wheel nose radius, and higher wheel speed all resulted in smoother surfaces.

Table 4 lists the changes in R_a values (in nm) as a result of changing a particular parameter. Note that the parametric coefficient represents half of the difference between the mean of the eight R_a values for the *high* state and the mean of the eight R_a values for the *low* state. For example, the wheel speed coefficient (Talysurf) was -36. Therefore, changing from a spindle speed of 30,000 rpm (the *low* state) to a spindle speed of 50,000 rpm (the *high* state) resulted in a 72 nm decrease in the R_a value.

Parameter	Decreases R_a	Increases R_a	R_a Change in nm (Zygo)	R_a Change in nm (Talysurf)
Coolant	water	oil	6	21
Wheel Speed	50k rpm	30k rpm	16	36
Part Material	silicon carbide	zerodur	9	11
Diamond Concentration	50	100	4	1
Part Speed	100 rpm	25 rpm	9	10
Part Diameter	6 mm	12 mm	11	9
Wheel Nose Radius	5 mm	1 mm	14	40
Diamond Particle Size	4-8 μ m	30-60 μ m	32	69
Dressing Technique	nib	stick	32	55
Depth of Cut	6 μ m	3 μ m	15	19
Crossfeed rate	0.15 mm/min	0.5 mm/min	22	43
Wheel Bond	resin	metal	9	11

Table 4: Relative effects of parametric changes on R_a values.

11.3 FUTURE WORK

The use of statistical design methodologies appears to have excellent potential for identifying the dominant parameters that control ductile regime grinding. The initial studies were limited in that a non-interactive linearization was assumed for the entire parameter space, surface finish measurements were used as a primary measure of the process, and reproducibility was not systematically checked. Surface finish, uncut chip profile geometry and critical cutting depths can be related by recently developed computer simulation techniques [3,4]; however, it is not yet known if surface finish itself is always an adequate measure of the extent of ductile-regime grinding conditions, especially when significant amounts of fracture damage are present in the finished surface.

While surface finish measurements will continue to be the primary means to assess the grinding process, these will be supplemented with surface fracture damage estimates using image analysis techniques developed in prior research work [5]. This procedure estimates the number density of microfracture (pit) sites on machined or ground surfaces using optical or SEM micrographs. By definition, the onset of the ductile-regime occurs at zero fracture density. Non-zero densities reflect the relative extent of ductile-regime conditions. A combination of surface finish and surface fracture damage estimates will be used to characterize ductile-regime conditions achieved for a given set of operating parameters.

It is clear from the results of the experiments reported here that diamond grit size, dressing technique and crossfeed rate have a significant effect on the process. Coolant chemistry, wheel bond and material selection are also expected to have large effects. The implicit assumption of non-interactive parameters, for example, that coolant chemistry or wheel bond will have exactly the same effect on silicon carbide (SiC) as on Zerodur, is obviously not realistic in many cases, and can mask the actual trends. Nevertheless, the preliminary results in Table 4 serve as a useful guide for further study.

Subsequent experiments will be designed to reduce the parameter space and expand the statistical design concepts underlying the experiments to allow interactive effects and experimental reproducibility to be systematically studied. An appropriate high/low range of conditions will be selected for each of the selected parameters. By appropriate design of the experimental test sequences, including certain repetitions and redundancy in the design matrix, the relative effect of each parameter, the interaction between parameters and the experimental reproducibility can be systematically determined with a minimum number of tests.

References

1. Bifano, Thomas G., "Ductile Regime Grinding of Brittle Materials," *PhD Dissertation*, North Carolina State University, Raleigh, NC, 1988.
2. Bifano, "Ductile Regime Grinding of Brittle Materials," *Precision Engineering Annual Report*, North Carolina State University, Raleigh, NC, Vol. V, 1987, p.254.
3. Fawcett, Steven C. and T. A. Dow, "Contour Grinding of Brittle Materials," *Precision Engineering Annual Report*, North Carolina State University, Raleigh, NC, Vol. VIII, 1990, pp. 225-237.
4. Fawcett, "Development and Implementation of a Grinding Technique for Precision Finishing of Brittle Materials," *PhD Dissertation*, North Carolina State University, Raleigh, NC, 1991.
5. Tidwell, R. Michael, "Ductile Regime Grinding of Germanium: Development of New Experimental and Analytical Analysis Methods," *MS Thesis*, North Carolina State University, Raleigh, NC, 1991.

12 MATERIAL FACTORS IN PRECISION GRINDING

Stanley M. Smith

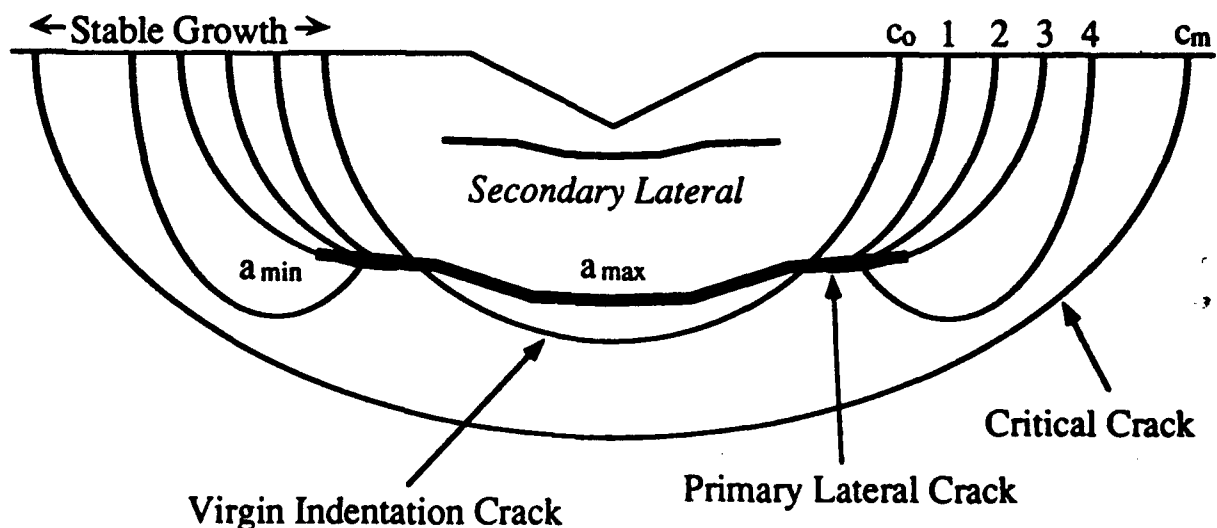
Graduate Student

Ronald O. Scattergood

Professor

Department of Materials Science and Engineering

Short-crack fracture toughness is a critical material property which influences the brittle to ductile transition and thus performance of brittle materials during precision grinding. The short crack fracture toughness can be determined by in situ measurement of indentation crack length as a function of applied stress. The in situ method has a basic advantage over other indentation methods because no assumptions regarding the indentation residual stress field are required. Verification tests on soda-lime glass indicate that crack shape strongly influences the results obtained from indentation tests. Radial cracks which intersect the surface were observed to be of quasielliptical shape rather than the semicircular shape that is usually assumed. Modification of the basic indentation fracture mechanics applicable to the in situ tests have been made for elliptical crack shapes. Incorporation of crack shape factors into the fracture mechanics analysis make possible accurate measurement of the short-crack fracture toughness and will provide important information regarding the fundamental mechanisms which control precision grinding performance.



12.1 INTRODUCTION

Prediction of the conditions which favor ductile mode grinding requires a fundamental understanding of wheel/workpiece interaction. Grinding performance depends on a combination of machine parameters and material properties. Properties of brittle materials are often size-scale dependent and/or influenced by environmental factors. Important properties such as fracture toughness and hardness can thus be altered by variation of grinding conditions. For different materials, very different grinding conditions may be required to obtain an acceptable surface finish. A global list of optimum machining parameters for an arbitrary workpiece material cannot therefore be constructed from exploratory grinding tests using a just a few materials. To obtain a more complete understanding of material removal processes, it is necessary to assess independently the variation of material properties with factors such as size-scale and environment. With this knowledge in hand, material selection can be combined with machine design to optimize precision grinding processes.

Fracture toughness values for short cracks are important in situations where small crack-size scales limit performance. Such is the case during precision grinding [1, 2], for which the critical factor is the initiation of cracks under a localized elastic-plastic deformation contact [3, 4]. A problem arises with many ceramic-based materials because the toughness is not constant, but instead increases with increasing crack-size scale. This behavior, known as the R-curve effect [5], is due to interaction of the crack with the microstructure [5-9]. In these types of materials, conventional measurements of toughness using long cracks will not give values appropriate for short-crack behavior.

Standard procedures for fracture toughness testing of ceramics are based upon the production of a precrack in a test specimen, followed by compliance or failure-stress measurements [10]. The precrack geometry must be well-defined so that stress intensity factors can be accurately evaluated. Straight, through-section cracks in bars or plates are preferred for most test configurations. However, it is difficult, if not impossible, to produce straight precracks having crack lengths below about 1 mm. Consequently, the standard long-crack fracture toughness test procedures are not adaptable to short cracks with lengths in the range of 50 μm or smaller.

Indentation methods using Vickers or Knoop hardness indenters have been used to generate short precracks for toughness testing. Indentation cracks of size as small as 20 - 50 μm can be produced for most ceramics and glasses. Either direct measurement of the crack length as a function of indentation load [11], subsequent measurement of an appropriate failure stress [12], or measurement of crack length during bend testing [13] can be used to estimate the toughness. Two important aspects complicate the use of indentation cracks. First, the stress intensity factor due to the indentation plastic zone must be incorporated into the analysis. This requires the introduction of

a residual-stress factor χ , which scales indentation load to the crack driving force. Second, the geometry of the indentation crack must be known. Half-penny crack shapes, in either median/radial [14] or Palmqvist geometries [15], are assumed in most analyses.

The present work deals with the measurement of short-crack toughness using indentation precracks. The methodology adopted requires *in situ* measurement of crack length during bend testing. The factor χ can, in principle, be independently determined as part of the test procedure for the specific material of interest rather than calibrated from separate measurements or model assumptions. Verification tests using soda-lime glass as the test material are reported here. Crack interaction and shape effects play a very significant, and rather unexpected role in the interpretation of the results.

12.2 FRACTURE TOUGHNESS MEASUREMENTS

12.2.1 Fracture Mechanics

Fracture toughness analyses for brittle materials generally assume that indentation precracks have the classic half-penny (semicircular) radial/median geometry [11, 12]. However, there is no compelling evidence that this is universal. Serial sectioning of indentation cracks often reveals a surface-localized, Palmqvist crack geometry [16]. In a recent review, Cook and Pharr [17] questioned the assumption of the general validity of half-penny radial/median geometry for Vickers indentation cracks in a variety of different ceramics and glasses.

Figure 1 shows a schematic of the radial/median and Palmqvist crack geometries produced by Vickers indentation. To avoid confusion with nomenclature, the term "radial" will be used here to denote a fully formed radial/median encompassing the indentation site. "Palmqvist" cracks are surface localized at opposite flanks of the indentation site. The crack shapes can be semicircular, but this need not be the case. In the following, the basic fracture mechanics equations for indentation-bend testing are developed.

For a radial crack generated by Vickers indentation, the stress intensity factor K for a crack length c , indentation load P , and applied stress σ is given by [18, 19]

$$K(c) = K_{\text{bend}} + K_{\text{residual}} = \psi_0 f_{\text{bend}} \sigma c^{1/2} + F_{\text{residual}} \frac{\chi P}{c^{3/2}} \quad (1)$$

The K_{bend} term corresponds to the stress intensity factor for an applied bend stress σ while the K_{residual} term results from the indentation residual-stress field due an indentation load P . Bending stress will be assumed constant over the crack profile. $\psi_0 (=1.29)$ [20] is a crack-shape factor for

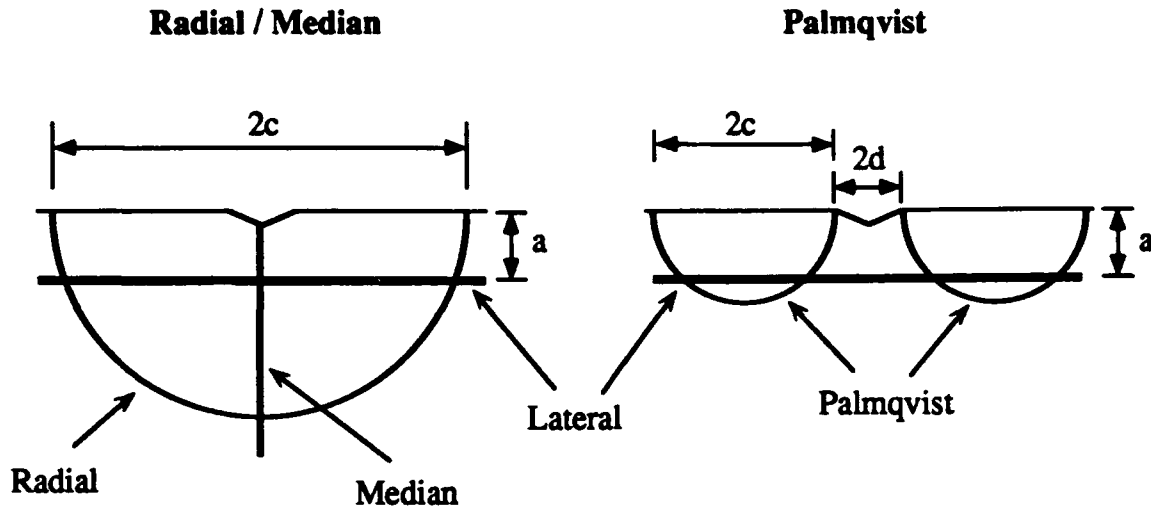


Figure 1: Schematic of indentation crack geometry.

semicircular cracks. χ is a residual-stress factor defined such that $\chi P = F/\pi^{3/2}$, where P is the indentation load and F denotes the magnitude of a point force that approximates the residual-stress driving force for an indentation crack. The factor $\pi^{3/2}$ has been included in $\chi P = F/\pi^{3/2}$ in order to recover the usual relationship between P and K_{residual} for embedded center-loaded penny cracks. χ cannot be determined *a priori*, but must be calibrated from experiments. f_{bend} and F_{residual} are correction factors accounting for departures from semicircular indentation crack shape.

The general form of the stress intensity factor $K(c)$ is given by Equation (1). The equilibrium length c of a crack at stress σ is given by

$$K(c) = \psi_0 f_{\text{bend}} \sigma c^{1/2} + F_{\text{residual}} \frac{\chi P}{c^{3/2}} = K_c \quad (2)$$

where K_c represents the fracture toughness of the material. If toughness is not constant (R-curve effect), K_c must be replaced by a toughness-curve relation $T(c)$ [5]. Due to the K_{residual} term in Equation (2), there will be stable crack growth prior to failure in a bend test whenever $\chi > 0$ [12].

A technique well-suited for determination of the short-crack fracture toughness involves *in situ* measurement of crack length as a function of applied stress, i.e., the *stable growth* of an indentation precrack is monitored, for example, using a bending stage on an optical microscope [13, 18]. The σ vs c data obtained establish the relation $\sigma(c)$ for fixed P . If crack shape (i.e. f_{bend} and F_{residual}) and χ are known, Equation (2) will give K_c directly for the available range of crack lengths. Alternatively, Equation (2) can be rearranged to give

$$\frac{f_{\text{bend}} \sigma c^2}{F_{\text{residual}}} = \frac{K_c c^{3/2}}{\psi_0 F_{\text{residual}}} - \frac{\chi P}{\psi_0} \quad (3).$$

The *in situ* σ vs c data can be analyzed by plotting $f_{\text{bend}} \sigma c^2 / F_{\text{residual}}$ vs $c^{3/2} / F_{\text{residual}}$. The least-squares slope and intercept allow K_c and χ to be independently determined. If the material shows R-curve behavior, then the slope is $T(c) / \psi_0$ where $T(c)$ denotes the toughness curve relation. In principal, measurement of the slope can be used to establish the form of the R-curve for crack sizes in the range commensurate with indentation precracks.

12.2.2 Experimental Procedure

Soda-lime glass was chosen as a test material for evaluating and verifying the *in situ* toughness determination technique. Soda-lime glass is free of microstructural constraints and has a constant toughness that has been characterized by other methods.

Soda-lime glass disks nominally 25 mm in diameter and 3.2 mm in thickness were stressed in biaxial flexure by a ring-on-three-ball loading fixture mounted on an inverted metallurgical optical microscope equipped with long-working-distance lenses. The radius of the circle upon which the three support balls were arranged was 10 mm. The loading ball was a 4 mm radius WC bearing ball precision ground to a 2 mm radius flat and countersunk to provide ring loading around its periphery. The support balls were WC bearing balls with a radius of 2 mm. The load was applied hydraulically through a set of three miniature bellows, the load being controlled by extension of a micrometer-driven piston into a reservoir of hydraulic fluid. Loads were monitored by a miniature strain-gauge load cell. A computer-based video image system was used to measure crack growth during loading. Figure 2 shows a schematic of the system. The ring-on-three-ball loading configuration results in constant stress within the periphery of the upper loading ring. Stresses were calculated using the elasticity equations given by Marshall [21].

The glass disks were annealed to remove any residual stresses before testing. Vickers indentation precracks were placed in the center of disks using indentation loads between 5 and 80 N and a dwell time of 10 s. The samples were indented using a hardness tester¹ in laboratory air and transferred to the loading stage in less than 1 min. To minimize environmental crack-growth effects during *in situ* tests, a dry nitrogen atmosphere surrounded the specimen in the loading stage. Stress was applied in a stepwise fashion, the stressing rate being approximately 1 MPa/s with hold periods of about 5 s at 2.5 - 5 MPa intervals. Video images of the cracks were acquired during the hold times. Crack length c is defined for radial and Palmqvist geometries in Figure 1. For each stress value, c was measured from the center of the indentation to each of the four crack

¹ Model 3212, Zwick Inc., East Windsor, CT.

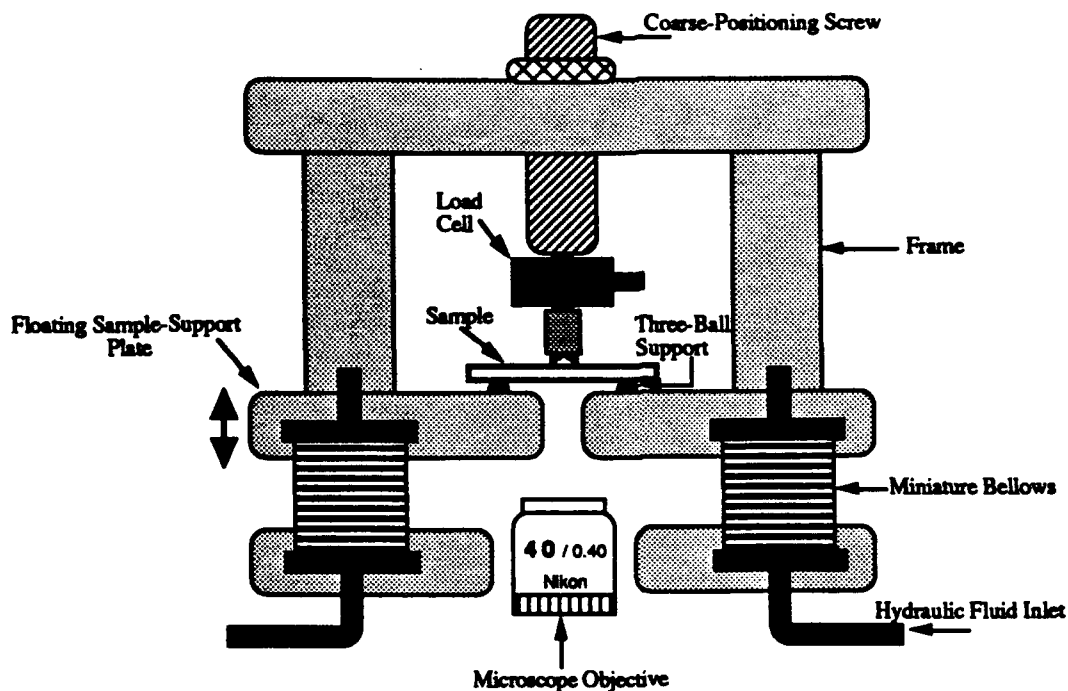


Figure 2: Schematic of the bending stage and microscope system used to obtain σ vs c data.

tips, and from the corners of the indentation to the four tips. Poorly defined cracks were rejected. Five separate runs were made at each indentation load.

Sub-surface crack shape profiles were observed by examining the fracture surfaces of indented soda-lime glass bars broken in four-point-bending. Intermediate crack shapes for crack lengths between the initial value c_0 and that at failure c_m were observed by introducing an aligned row of test indentations at load P . A higher-load indentation was made to serve as the strength controlling flaw at the end of the aligned row. The critical crack from the large indentation passed through the center of the smaller indentations, thus providing a cross-section. Depending on the load value used for the large indentation, different degrees of subcritical crack growth can be generated for the test indentations.

12.3. RESULTS AND DISCUSSION

12.3.1 Experimental Results

Figure 3 shows typical σ vs c data plotted according to Equation (3) using the semicircular radial crack shape assumption, i.e., $f_{\text{bend}} = F_{\text{residual}} = 1$ (the axes have been normalized by P to display a series of different runs). The plots in Figure 3 are linear except for the first few points, the start-up transients being attributed to environmental crack growth that occurred immediately after indentation in air, prior to *in situ* bend testing. Figure 4 shows values of K_c and χ vs P derived from the σ vs c data using the radial semicircular crack shape assumption. K_c and χ values were

determined from the least-squares slopes and intercepts of the linear portion of σc^2 vs $c^{3/2}$ plots. Both K_c and χ increase with P over the range of indentation loads investigated. This result is quite unexpected since soda-lime glass should not show R-curve behavior. Furthermore, the values of fracture toughness at all indentation loads shown in Figure 4 are significantly higher than the value $K_c = 0.75 \text{ MPa}\sqrt{\text{m}}$ usually quoted for soda-lime glass. There is no physical basis for the apparent R-curve effect. The results shown for soda-lime glass indicate that the assumption of half-penny radial cracks is not valid.

12.3.2 Crack Geometry

The dependence of the actual crack shape on indentation load and bend stress was ascertained from visual observations on cross-sections of fracture surfaces [19]. Departure from half-penny crack shape was found for all of the indentation loads used. This was a result of interaction of the radial crack extending outward from the indentation site with the lateral crack. The key feature is that the radial crack is constrained in the depth dimension. Based on a large number of cross-section observations, the actual crack geometry is summarized schematically in Figure 5.

A lateral crack intersecting the radial crack was observed at all indentation loads. The lateral crack usually deflected upward away from the indentation site, as shown in Figure 5. Downward deflection was occasionally observed, as reported in earlier investigations [14, 22]. Because of these deflections, lateral-crack depth varies from a_{max} to a_{min} . Upon subsequent stressing in the *in situ* tests, the radial indentation crack undergoes stable growth through positions 1-4 up to a critical size c_m at fracture. Because of the constraint imposed by the primary lateral crack, the radial crack shape is altered, with the lower portion being truncated by the lateral. When the radial reaches the outer extent of the lateral, it can "wrap around" and bulge out as shown at location 4 in Figure 5. With further extension, the two wrapped ends of the radial can rejoin, subsequently forming a critical crack free of lateral constraint. The important point for analysis of σ vs c data is that crack shape departs substantially from a true half-penny over the entire growth-to-failure path.

To include crack-shape effects into the analysis, a physically realistic model for the constraint imposed by the primary lateral cracks must be developed. The depth a of the lateral crack will be assumed constant for given P . The crack-shape constraint effect will be incorporated into a one-parameter, elliptic-crack model where the ellipticity is governed by the depth of the lateral. Two limiting crack geometries are possible. The first is a modified radial crack system where the crack shape is elliptical rather than half penny. The second is a modified Palmqvist crack system, where again the crack shape is elliptical rather than half-penny. While radial crack geometry is the physically realistic choice based on the fracture cross-section observations, Palmqvist crack geometry will also be treated in light of the results reported in [17]. The correction factors for elliptical cracks, $f_{\text{bend}}(e)$ and $g_{\text{residual}}(e)$, are displayed in Figure 6(a), while $f_{\text{residual}} = f_{\text{R}_{\text{residual}}}(e)$ for radial cracks and $f_{\text{residual}} = f_{\text{PQ}_{\text{residual}}}(e, d)$ for Palmqvist cracks are shown in figures 6(a) and

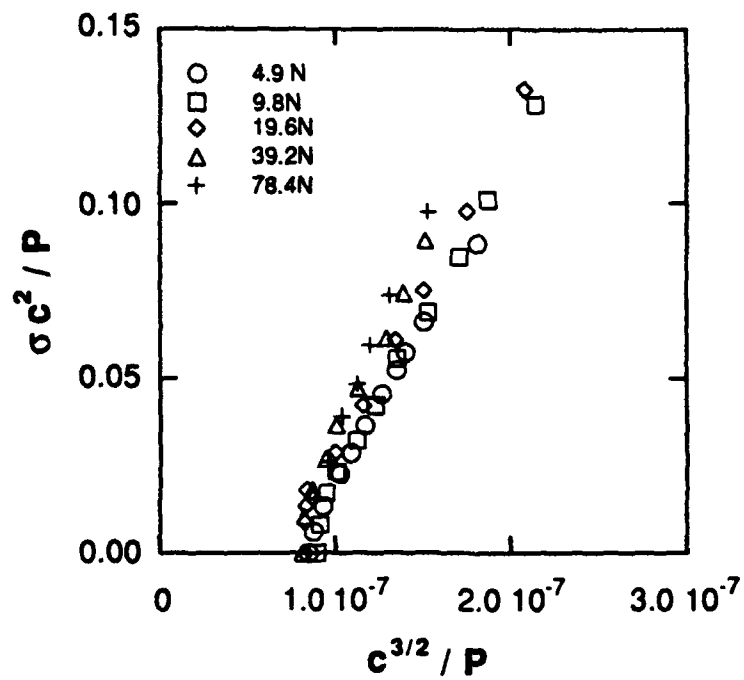


Figure 3: $\sigma c^2/P$ vs $c^{3/2}/P$ for the P values indicated.

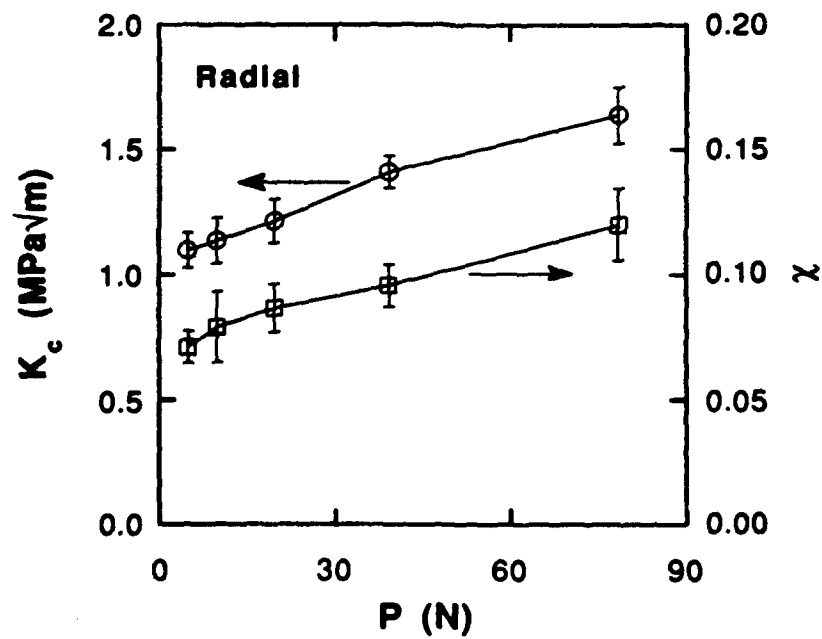


Figure 4: K_c and χ values derived using half-penny radial crack geometry. Error bars represent one standard deviation.

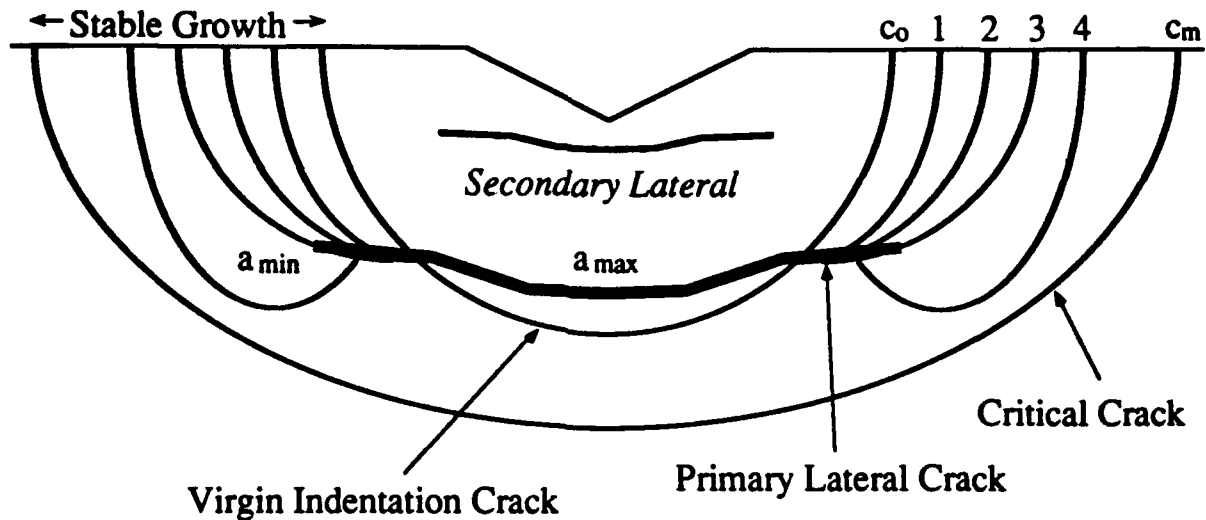


Figure 5: Schematic representation of the observed sub-surface indentation crack shapes.

6(b), respectively. Note that the latter depends on indentation diagonal length as well as ellipticity. $g_{\text{residual}}(e)$ is a correction factor accounting for free surface effects ($F_{\text{residual}} = f_{\text{residual}} g_{\text{residual}}$). The calculational procedures employed to evaluate these factors are given in [19]. The analysis appropriate to the radial or Palmqvist crack geometries differs only in the choice of the crack length c and the correction factor F_{residual} to be used in Equation (3).

12.3.3 Crack Ellipticity

To apply Equation (3) to σ vs c data, the crack ellipticity $e = c/a$ must be known. Theory asserts that lateral crack depth a should scale with the indentation plastic zone size [17], hence, with $P^{1/2}$. The relation for crack ellipticity is then

$$e = \frac{c}{a} = \frac{c}{\lambda P^{1/2}} \quad (4)$$

where the constant $\lambda = a/P^{1/2}$ provides a convenient means to introduce ellipticity into Equation (3). In this formulation, λ represents the adjustable parameter in the elliptic-crack model. Lateral crack depths were measured experimentally on cross-section fracture surfaces. Since the lateral depth varied as shown in Figure 5, minimum and maximum values of the depth were determined at each P using an average of ten measurements in each case. Plots of a vs $P^{1/2}$ are shown in Figure 7, from which values of $\lambda_{\text{min}} = 1.2 \times 10^{-5} \text{ m/N}^{1/2}$ and $\lambda_{\text{max}} = 1.8 \times 10^{-5} \text{ m/N}^{1/2}$ are obtained for a_{min} and a_{max} , respectively.

Figure 8 shows typical σ vs c data plotted according to Equation (3) for elliptical radial crack geometry (the axes have been normalized by P to display a series of different runs). The curves are

linear beyond the initial start-up transient. Similar plots were obtained for the case of Palmqvist geometry. Values of K_C and χ were determined from least-square slopes and intercepts fitted to the linear portion of the plots.

Figure 9 shows the dependence of K_C and χ on P for a range of λ values assuming elliptical radial crack geometry. For large values of λ , there is no lateral crack constraint and the radial crack shapes are half-penny. As λ decreases, ellipticity increases and the crack-shape correction terms

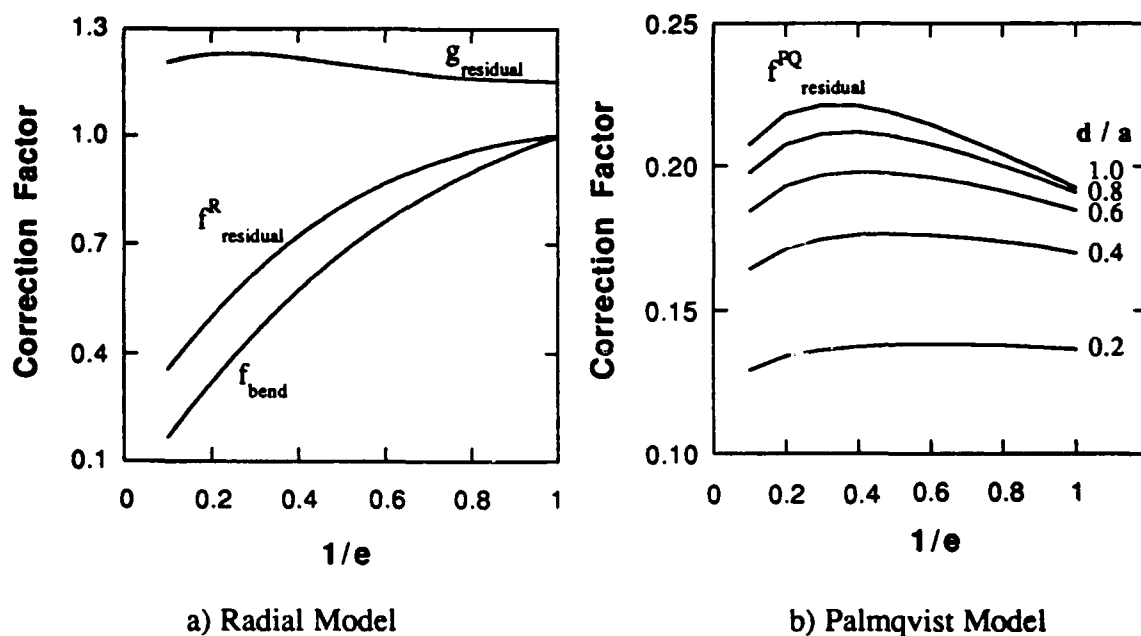


Figure 6: Ellipticity correction factors.

become more important. The apparent R-curve effect diminishes and the magnitude of the toughness decreases as λ is decreased. Figure 9 also shows the bounds (solid points) for the experimentally determined λ_{\min} and λ_{\max} . At $\lambda_{\max} = 1.8 \times 10^{-5} \text{ m/N}^{1/2}$, both K_C and χ are nominally constant with P , the *average value* and standard deviation for all tests being $K_C = 0.54 \pm 0.07 \text{ MPa}\sqrt{\text{m}}$ and $\chi = 0.029 \pm 0.007$. In view of the fact that $\lambda < \lambda_{\max}$ produces unrealistically low values of K_C for soda-lime glass, λ_{\max} will be chosen as the "best-fit" value.

Figure 10 shows the dependence of K_C and χ on P for a range of λ values assuming elliptical Palmqvist crack geometry. The apparent R-curve effect again diminishes and the magnitude of the toughness decreases as λ is decreased. However, compared to the radial crack model in Figure 9, the trends and overall changes in the data are much less pronounced. The bounds for the experimentally determined λ_{\min} and λ_{\max} are shown in the figure (solid points). The "best-fit" for

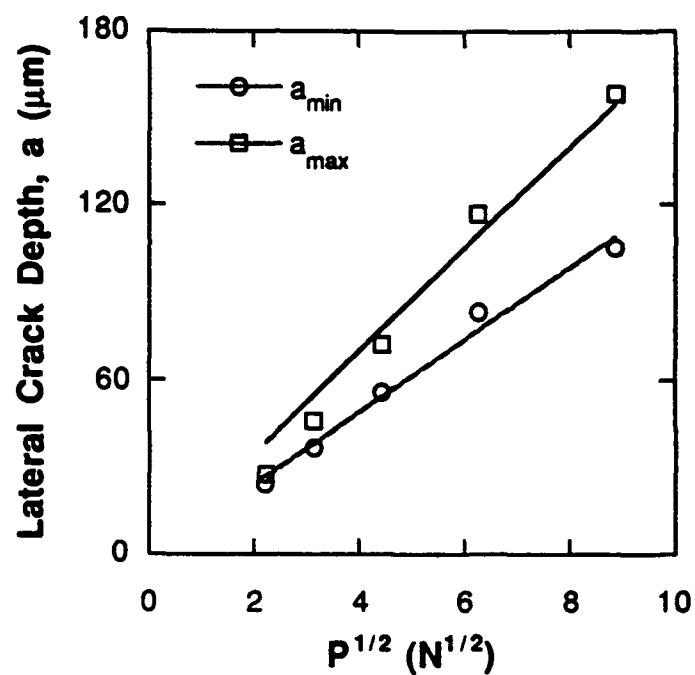


Figure 7: a_{max} and a_{min} vs $P^{1/2}$.

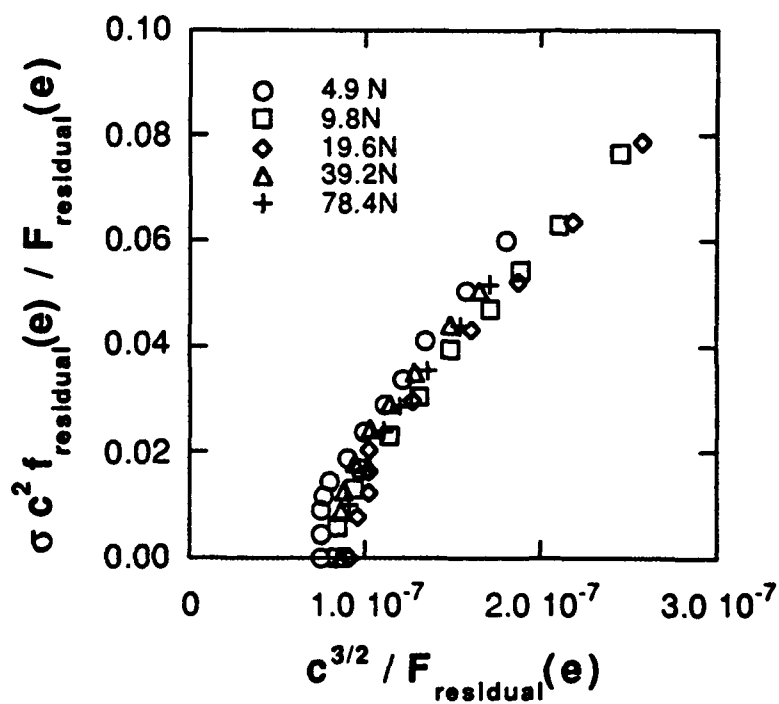


Figure 8: $\sigma c^2 f_{residual} / F_{residual} P$ vs $c^{3/2} / F_{residual} P$ for the P values indicated.

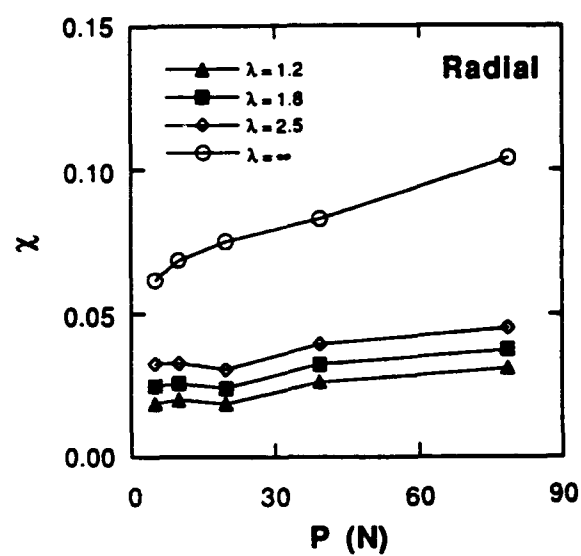
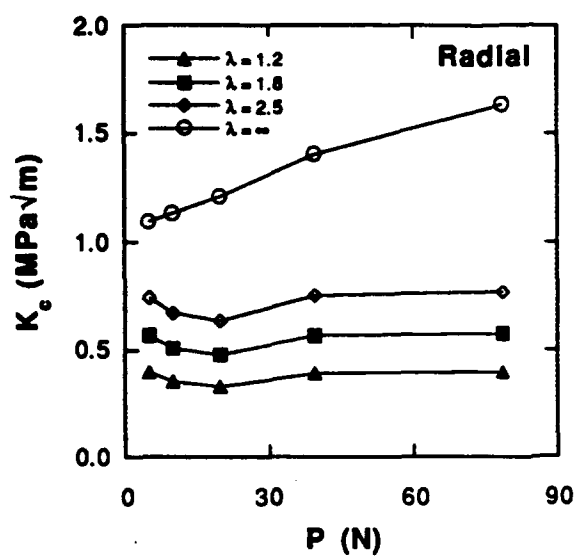
the Palmqvist model, in the sense of closest agreement with the radial-crack model, is obtained for $\lambda_{\min} = 1.2 \times 10^{-5} \text{ m/N}^{1/2}$ where $K_C = 0.55 \pm 0.07 \text{ MPa}\sqrt{\text{m}}$ and $\chi = 0.032 \pm 0.006$. Different choices of λ can be expected, because constrained radial cracks possess more ellipticity than do constrained Palmqvist cracks at fixed lateral-crack depth.

Figure 11 summarizes the classical half-penny crack predictions with those determined from the radial-crack and Palmqvist-crack models using the values of λ discussed above. The half-penny results shown do not contain the free-surface correction g_{residual} included for the half-penny limit $\lambda = \infty$ in Figures 9 and 10. It is clear that g_{residual} introduces minor corrections. An essential point to note is the fact that the elliptic-crack models incorporate correction factors that account for the changes in crack shape occurring during *in situ* bend tests. The *magnitude* of these corrections reduce K_C values, while their *change* with P eliminates the apparent R-curve effect. The latter is primarily a result of the lack of self similarity of the constrained crack shapes with P (self similarity being implicit in half-penny crack models).

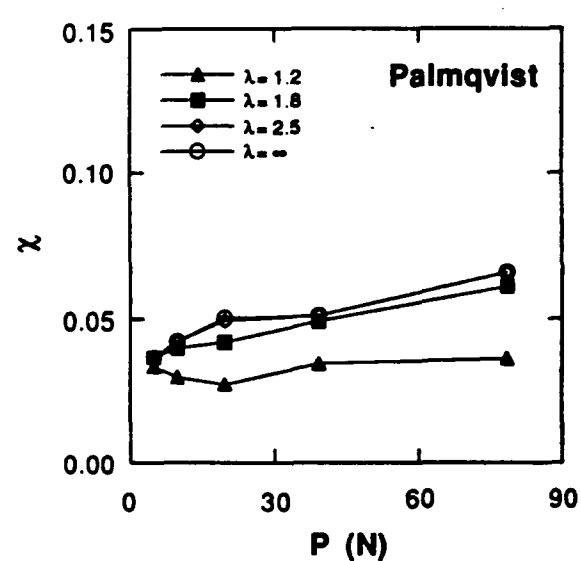
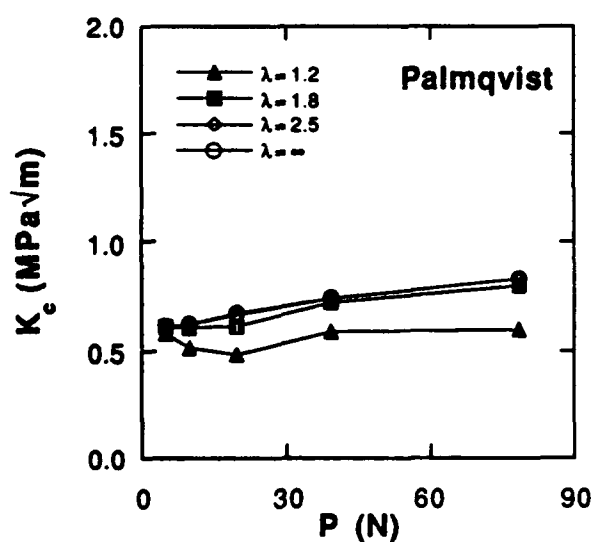
12.3.4 Discussion

The average value $K_C = 0.54 \pm 0.07 \text{ MPa}\sqrt{\text{m}}$ derived here is less than the widely quoted inert-limit $K_C = 0.75 \pm 0.05 \text{ MPa}\sqrt{\text{m}}$ for soda-lime glass measured using a fast-break DCB technique [23]. Environmental effects can reduce the apparent toughness. An environmental correction for elliptical cracks can be derived from simulation results [19]. This adjusts $K_C = 0.54 \text{ MPa}\sqrt{\text{m}}$ to an "inert limit" value $K_C = 0.65 \text{ MPa}\sqrt{\text{m}}$. In view of the potential difficulties encountered in testing, particularly when environmental effects, crack-shape effects and model assumptions are involved, the agreement between the *in situ* tests and the accepted value $K_C = 0.75 \text{ MPa}\sqrt{\text{m}}$ appears reasonable.

The results obtained in this investigation provide an *independent* measurement of the residual stress factor χ , a feature not possible with other test methods. Over the range of indentation loads used, the average value $\chi = 0.029 \pm 0.007$ was obtained. No decrease in χ with increasing indentation load could be found. Evidently, the decrease in χ due to relaxation effects associated with lateral crack growth [24] is not significant for the range of indentation loads $5 \text{ N} \leq P \leq 80 \text{ N}$. It must be recognized that the value of χ obtained here is for post-indentation crack growth. The operative value of χ may be different for the *formation* of virgin indentation cracks. Lawn et. al. [14] derived the value $\chi = 0.049$ from *in situ* measurements of the formation of indentation cracks prior to lateral-crack formation, suggesting that there may be some relaxation of the residual stresses upon formation of the laterals when the indenter load is removed.



(a) (b)
Figure 9: Radial elliptic-crack model. (a) K_c vs P and (b) χ vs P for the λ values indicated (λ in units of $10^{-5} \text{ m/N}^{1/2}$).



(a) (b)
Figure 10: Palmqvist elliptic-crack model. (a) K_c vs P and (b) χ vs P for the λ values indicated (λ in units of $10^{-5} \text{ m/N}^{1/2}$).

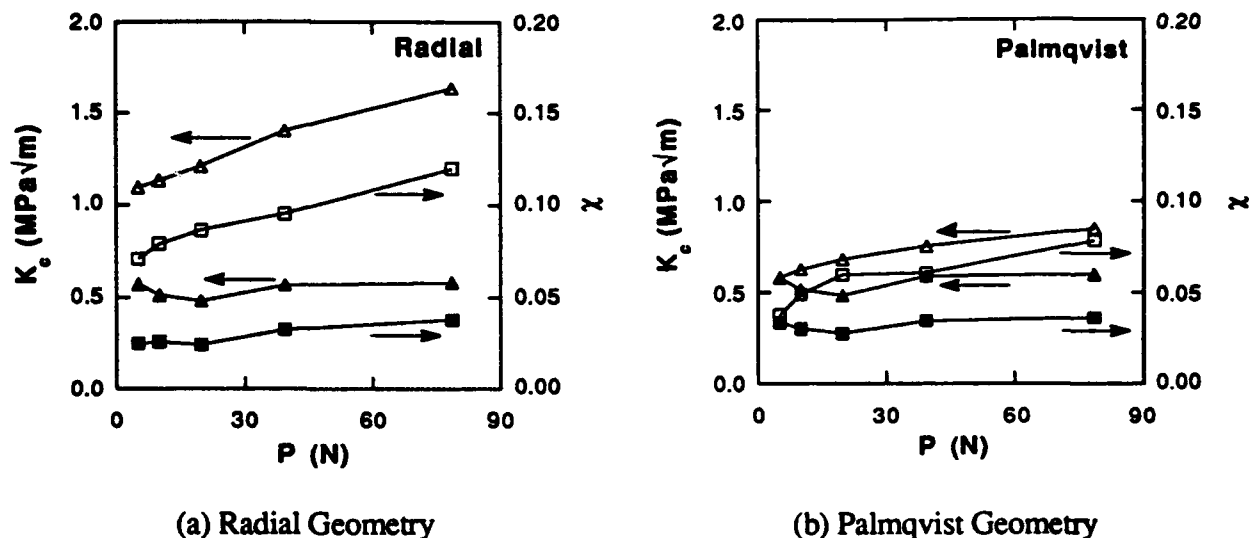


Figure 11: Predictions from the classical half-penny crack model (open points) and the best-fit elliptic-crack models (solid points).

12.4 SUMMARY AND CONCLUSIONS

Departures of the crack geometry from half-penny shape introduce significant modifications into the analysis of σ vs c *in situ* test data. In soda-lime glass, lateral cracks interact with and constrain the virgin indentation cracks so that, upon subsequent stable growth under applied stress, the cracks deviate markedly from half-penny geometry. The deviations can be described by a one-parameter model where the crack shape is assumed to be elliptical, the minor axis dimension being constrained by the depth of the lateral crack. Experimental measurements were made to establish the range of lateral crack depths. The *in situ* test results can be rationalized on the basis of an elliptic-crack model, using either radial-crack or Palmqvist-crack geometries. Based on the actual crack shapes observed in soda-lime glass, the radial-crack model is more realistic. Compared to the classical half-penny crack geometry, the introduction of ellipticity eliminates an apparent R-curve effect and reduces toughness values to the range expected for soda-lime glass. The average values $K_c = 0.54 \pm 0.07$ MPa√m and $\chi = 0.029 \pm 0.007$ were independently determined for soda-lime glass in a dry nitrogen environment. χ was found to be nominally independent of indentation load P over the range $5\text{N} \leq P \leq 80\text{N}$.

That the average toughness $K_c = 0.54$ MPa√m measured here is less than the "standard" value $K_c = 0.75$ MPa√m for soda-lime glass is most likely a consequence of the assumption that the crack shapes are truly elliptical. The actual crack-shape profiles shown in Figure 5 are best characterized as rectangular with an evolving curved "bulge" at the tip. *In situ* measurements capture much of

this bulge evolution during the test. The actual crack shapes are less constrained than elliptical shapes, especially near the crack tip. The real shape-correction factors would therefore be larger than for ellipses and the corrected toughness values would increase. Environmental effects also lead to smaller apparent toughnesses. Interestingly, the elliptical or quasi-elliptical shapes, constrained in the depth dimension by the laterals, have the feature that ψ decreases as crack length increases. This leads to a stabilization of crack growth prior to failure, above and beyond that due to the residual stress factor χ . An elliptical crack must in fact escape from the constraining lateral before failure (unstable crack growth) can occur.

The *in situ* method for measuring toughness using the stable growth regime of indentation precracks is attractive in that only the crack-shape need be known a priori. Other indentation fracture toughness methods require a value of the residual stress factor χ , a quantity very difficult to determine without further assumptions. The successful application of the *in situ* technique to other ceramics, and possibly to R-curve determination, will require a detailed characterization of crack shapes.

References

1. Bifano, T.G., T.A. Dow, and R.O. Scattergood, "Ductile-Regime Grinding of Brittle Materials: Experimental Results and the Development of a Model", *Proc. SPIE-Int. Soc. Opt. Eng*, **966** p. 108 (1988).
2. Marshall, D.B., B.R. Lawn, and R.F. Cook, "Microstructural Effects on Grinding of Alumina and Glass-Ceramics", *J. Am. Cer. Soc.*, **70**(6) C-139-40 (1987).
3. Hagan, J.T., "Micromechanics of Crack Nucleation During Indentations", *J. Mat. Sci* , **14**: p. 2975-80 (1979).
4. Chiang, S.S., D.B. Marshall, and A.G. Evans, "The Response of Solids to Elastic/Plastic Indentation II. Fracture Initiation", *J. App. Phy.*, **53**(1) 312-17 (1982).
5. Cook, R.F., B.R. Lawn, and C.J. Fairbanks, "Microstructure-Strength Properties in Ceramics: I, Effect of Crack Size on Toughness", *J. Amer. Cer. Soc.*, **68**(11) 604-615 (1985).
6. Evans, A.G. and R.M. Cannon, "Overview No. 48: Toughening of Brittle Solids by Martensitic Transformations", *Acta Met.*, **34**(5) 761 (1986).
7. Marshall, D.B. and A.G. Evans, "Failure Mechanisms in Ceramic-Fiber/Ceramic-Matrix Composites", *J. Amer. Cer. Soc* , **68**(5) 225-31 (1985).
8. Becher, P.F., C. Hsueh, P. Angelini, and T.N. Tiegs, "Toughening Behavior in Whisker-Reinforced Ceramic Matrix Composites", *J. Am. Cer. Soc.*, **71**(12) 1050-61 (1988).
9. Swanson, P.L., C.J. Fairbanks, B.R. Lawn, Y. Mai, and B.J. Hockey, "Crack-Interface Grain Bridging as a Fracture Resistance Mechanism in Ceramics: I, Experimental Study on Alumina", *J. Am. Cer. Soc.*, **70**(4) 279-89 (1987).
10. Anderson, R.M., "Testing Advanced Ceramics", *Adv. Mat. and Proc*, 31-36 (1989).
11. Antis, G.R., P. Chantikul, B.R. Lawn, and D.B. Marshall, "A Critical Evaluation of Indentation Techniques for Measuring Fracture Toughness: I, Direct Crack Measurements", *J. Am. Cer. Soc.*, **64**(9) 533-38 (1981).

12. Chantikul, P., G.R. Antis, B.R. Lawn, and D.B. Marshall, "A Critical Evaluation of Indentation Techniques for Measuring Fracture Toughness: II, Strength Method", *J. Am. Cer. Soc.*, **64**(9) 539-43 (1981).
13. Marshall, D.B., "Controlled Flaws in Ceramics: A Comparison of Knoop and Vickers Indentation", *J. Am. Cer. Soc.*, **66**(2) 127-31 (1983).
14. Lawn, B.R., A.G. Evans, and D.B. Marshall, Elastic/Plastic Indentation Damage in Ceramics: The Median/Radial Crack System", *J. Am. Ceram. Soc.*, **63**(9-10) 574 (1980).
15. Shetty, D.K., I.G. Wright, P.N. Mincer, and A.H. Clauer, "Indentation Fracture of WC-Co Cermets", *J. Mat. Sci.*, **20** 1873-82 (1985).
16. Shetty, D.K., A.R. Rosenfield, and W.H. Duckworth, "Indenter Flaw Geometry and Fracture Toughness Estimates for a Glass-Ceramic", *J. Am. Cer. Soc.*, **68**(10) C-282-84 (1985).
17. Cook, R.F. and G.M. Pharr, "Direct Observation and Analysis of Indentation Cracking in Glasses and Ceramics", *J. Am. Cer. Soc.*, **73**(4) 787-817 (1990).
18. Marshall, D.B., B.R. Lawn, and P. Chantikul, "Residual Stress Effects in Sharp Contact Cracking Part 2 Strength Degradation", *J. Mat. Sci.*, **14** 2225-35 (1979).
19. Smith, S.M. and R.O. Scattergood, "Crack Shape Effects for Indentation Fracture Toughness Measurements", Submitted to *J. Am. Ceram. Soc.*, 1991.
20. Oore, M. and D.J. Burns, "Estimation of Stress Intensity Factors for Embedded Irregular Cracks Subjected to Arbitrary Normal Stress Fields", *Transactions of the ASME*, **102** 202-11 (1980).
21. Marshall, D.B., "An Improved Biaxial Flexure Test for Ceramics", *Cer. Bull.*, **59**(5) 551-53 (1980).
22. Marshall, D.B. and B.R. Lawn, "Residual Stress Effects in Sharp Contact Cracking: Part 1 Indentation Fracture Mechanics", *J. Mat. Sci.*, **14** 2001-12 (1979).
23. Wiederhorn, S.M., "Fracture Surface Energy of Glass", *J. Am. Ceram. Soc.*, **52**(2) 99-105 (1969).
24. Cook, R.F. and D.H. Roach, "The Effect of Lateral Crack Growth on the Strength of Contact Flaws in Brittle Materials", *J. Mater. Res.*, **1**(4) 589-600 (1986).

13 NANOFABRICATION

Robert D. Day

Research Scientist

Graduate Student

Los Alamos National Laboratory¹

Materials Science & Engineering

Phillip E. Russell

Professor

Materials Science & Engineering

13.1 INTRODUCTION

Recent advances in methods for imaging and manipulating materials at the nanometer scale make it possible to consider a broad array of new approaches for fabrication of ultrasmall devices. A project is underway at Los Alamos National Laboratory to construct a Ultrahigh Vacuum (UHV) apparatus for investigating and developing new nanofabrication techniques. Coupling carefully chosen experiments with atomistic simulations will allow exploration of the limits of size and control achievable with techniques such as atomic scale cutting, ion beam milling and Taylor cone extrusion. Nanofabrication technology will have immense industrial importance by facilitating the next stage of miniaturization of mechanical, electronic, optical and chemical devices.

13.2 BACKGROUND AND SIGNIFICANCE

Nanofabrication, the formation, manipulation and characterization of structures on the nanometer scale, seemed all but impossible until ten years ago. In 1982 Binnig et al. [1] announced the invention of the scanning tunneling microscope (STM). Since then, advances in atomic scale imaging and manipulation technologies have occurred almost yearly. The Atomic Force Microscope (AFM), which is based upon forces sensed by the tip, can be used to image nonconducting materials and measure microscopic frictional effects. With an AFM, one can scrape or indent a surface while measuring the resisting force, and subsequently image the modification with the same tip. A different and somewhat more destructive nanoscale device is the liquid metal ion source used for Focused Ion Beam (FIB) systems. State-of-the-art, Ga liquid FIB's can now potentially cut channels as small as 50 nm wide, and image square millimeter areas with submonolayer destruction. Because of the extremely small registry errors of the FIB, its

¹ Additional Collaborators at Los Alamos National Laboratory

Marilyn Hawley, Joel Kress, Bruce Lamartine and Art Voter

combination with AFM/STM offers the opportunity for recurring passes of a machine tip to produce novel atomic scale structures.

With these developments, nanofabrication has become the new frontier in the miniaturization of mechanical, electronic, optical, chemical, and even biological devices. There is no longer a question that fabrication of atomic scale features will be possible. The new question becomes: just what are the limits?

A very attractive feature of nanofabrication technology is the tight connection that can be made between experiment and the modeling. The minimum scale on which materials can be characterized and manipulated has been decreasing for centuries, while the system size that can be modeled atomistically in a computer has been increasing since the invention of Molecular Dynamics (MD) in the 1950's. At the nanometer scale, the size scales accessible to these two converging technologies overlap for the first time.

13.3. APPROACH

State-of-the-art experiments will be coupled with large-scale atomistic modeling. The theoretical and experimental tools described here will allow for an exploration of the limits of nanofabrication techniques.

13.3.1 Simulation Methods

The basic tools in the modeling effort are MD, in which classical equations of motion are integrated for a system of atoms. The system size that can be treated by MD is limited by the speed of the computer, while the accuracy of the predicted properties is limited by the quality of the interatomic potential. Los Alamos has an MD program capable of modeling one to ten million atoms on the CM2 connection machine. (This code is presently designed for 2-D, but will be adapted to 3-D). In many cases, questions can be addressed without the use of large-scale MD, by applying molecular statics or MD to a properly chosen representative system. For events requiring longer time scales (e.g., diffusion backfilling of a milled feature) the overlayer dynamics method can be employed. Regarding the accuracy of simulations, LANL has state-of-the-art interatomic potentials for carbon, silicon, and many metals, and expertise in the development and use of other advanced forms of many-body potentials.

13.3.2 Experimental Apparatus

A system is being constructed to take special advantage of the control that Ultrahigh Vacuum (UHV) permits for nanofabrication. A target base pressure of 2×10^{-11} Torr will be sought to maintain the highest purity of substrate surfaces and formed nanostructures. The key techniques

available will be STM, AFM, FIB's of several species, ultrapure vapor deposition sources allowing both Physical Vapor Deposition and Chemical Vapor Deposition, and several surface analysis techniques (all available by appropriate use of a suitable particle energy analyzer). With the use of rapid sample transfer in vacuo, quick feedback of the outcome of each nanofabrication experiment can be obtained.

13.3.3 Proposed Monolayer Removal Experiments

One of the proposed experimental topics for the nanofabrication system will be that of mechanically removing a monolayer of material from a suitable substrate. This will be accomplished by using an AFM or STM stylus. A variety of system effects must be investigated to adequately address this topic. Some of these topics relate to the stylus' chemical and geometrical properties. Others include rate, force, and cutting scheme effects. Finally, the effects of the cutting environment and the surface's material properties must be investigated. Manipulation of a few atoms, or even a few thousand atoms, reveals phenomena of primary importance that are considered to be second or third order for larger size scales. A prime example is surface diffusion. Depending on factors such as the choice of material, temperature, or orientation, a monolayer deep channel may be rapidly refilled by diffusing atoms. Molecular statics calculations will be used to determine relevant diffusional barrier heights, and overlayer dynamics techniques can be applied as necessary. These results will increase the understanding of how to control this diffusion. The limits of the maximum sharpness attainable for the cutting stylus will also be investigated. It is anticipated that the stylus will be sharpened with a FIB. Simulations will be useful in determining stylus geometries expected to be most stable against diffusional blunting. Large scale MD simulations will be employed to model the actual cutting process.

By coupling experiments with calculations, a more thorough understanding of the mechanical material removal process at the nanometer level will be gained at a rate higher than could be achieved through experimentation alone.

13.4 CURRENT STATUS

A FIB system has been delivered which is touted as having a 50 nm spot size, as well as a five axis UHV compatible hot stage. Ion and turbomolecular pumps have been ordered and the designs for most of hardware which will go into the system are being completed. A commercial UHV compatible STM with a lithography software package will be purchased. The Nanoscope I which was "computerized" at the PEC will also be used to perform some of the STM and AFM related experiments.

13.5 SUMMARY

A Nanofabrication System is being constructed at LANL which will take advantage of state-of-the-art AFM, STM, FIB, and a variety of surface analysis capabilities. In conjunction with this hardware, state-of-the-art computer modeling will be used to better define our experiments and to gain a better understanding of material modification processes which occur at the nanometer level.

References

- [1] G. Binning, H. Rohrer, Ch. Gerber, and E. Weibel, "Surface Studies by STM", *Physical Review Letters*, Vol. 49, No. 1, 1982, p. 57.

14 MODEL REFERENCE ADAPTIVE CONTROL OF DUAL-MODE MICRO/MACRO DYNAMICS OF BALL SCREWS FOR NANOMETER MOTION

Peter I. Hubbel

Graduate Student

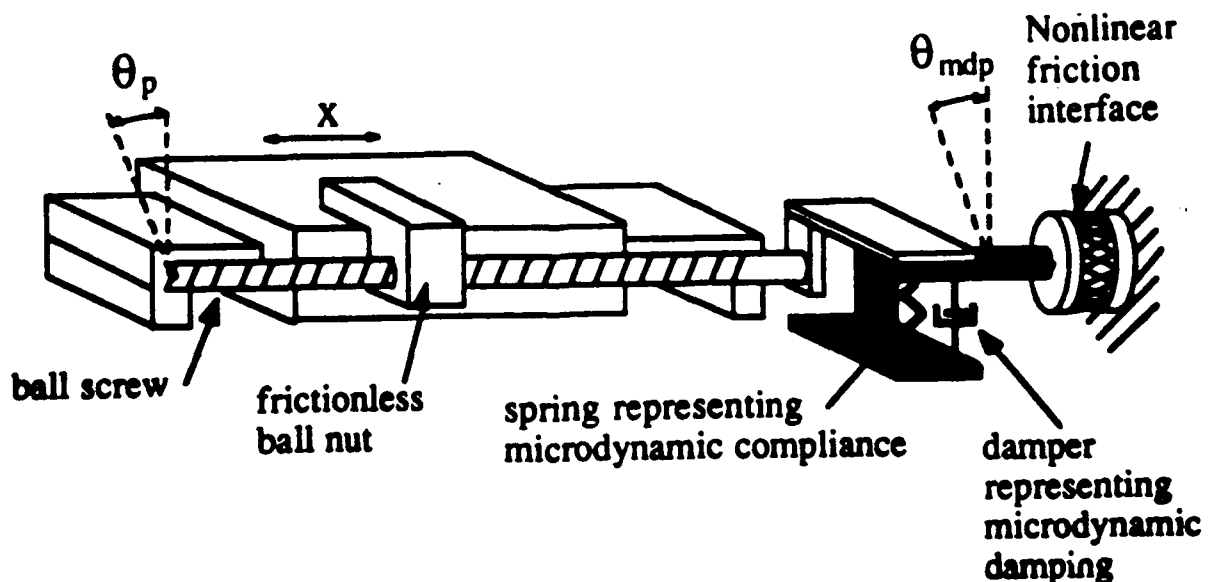
Department of Electrical and Computer Engineering

Paul I. Ro

Assistant Professor

Department of Mechanical and Aerospace Engineering

A desire to improve the positioning accuracy of ball screws prompted an investigation into the dynamics of nanometer motion. Characterization of the ball screw indicated that nanometer motion is possible prior to friction breakaway via elastic deformation of the frictional contacts while macroscopic motion requires slipping across the friction interfaces. The observed dynamics are nonlinear, and consequently result in inconsistent and unpredictable closed-loop response while under PI position control. The ball-screw can be modeled in two stages, the microdynamic stage includes "elastic" friction while the macrodynamic stage incorporates kinetic (sliding) friction. A two-stage model reference adaptive control (MRAC) strategy is adopted and a Lyapunov design technique is applied to derive the adaptive laws. Experimental results obtained via a DSP implementation of the adaptive controller indicate that each stage of the adaptive control performs well within the respective dynamic regions, but performance deteriorates as either controller is operated near the boundary of the regions.



14.1 INTRODUCTION

The increasing demands that precision engineering applications place on positioning systems has prompted an investigation into the nanometer positioning capabilities of ball screws. Though ball screws have been used for many years, they traditionally were not applied to problems requiring sub-micrometer accuracy.

As in many mechanical systems, nonlinear friction in the ball screw complicates the control design process and can significantly reduce the positioning accuracy. Various position control techniques have been presented which deal with macroscopic stick/slip behavior [1-3]. However, the results presented here indicate that attention must also be paid to the peculiar nonlinear dynamics (microdynamics) which dominate sub-micrometer motion.

Characterization of the ball-screw for nanometer and long range motion indicates that the dynamics can be separated into two stages. The first (microdynamic) stage is dominated by elastic deformation of the frictional contacts between the balls and the grooves in the nut and screw [4]. An earlier study of metallic friction interfaces revealed similar microscopic elastic behavior [5]. The second (macrodynamic) stage is characterized by a classical kinetic (sliding) friction model. Similar dynamics have also been observed in a ball-bearing slideway [6].

A new piecewise linear model is developed to reproduce the general behavior of the ball-screw and serve as a testbed for control schemes. The model includes two stages, the first is the microdynamic stage and is applicable to sub-micrometer motion while the second stage is macrodynamic and intended to reproduce larger range dynamics. The microdynamic stage includes an elastic friction model while the macrodynamic stage uses a classical kinetic friction model.

Because each of the linear model segments include unknown parameters, and preliminary closed loop experiments indicated that a PI control law resulted in an unpredictable and inconsistent closed loop performance, an adaptive control approach is selected. The adaptive control scheme includes two stages. One was designed using the microdynamic plant model and was intended for sub-micrometer motion. The other was based upon the macrodynamic plant model and was intended for larger range motion.

Model reference adaptive control (MRAC) is selected because it allows for easy shaping of the closed-loop response. A Lyapunov technique is used to derive the adaptive laws to guarantee convergence of the plant and reference model [7, 8]. Whereas some Lyapunov techniques result in adaptive laws which rely solely on velocity error, the implementation described here uses position and velocity error [9, 10].

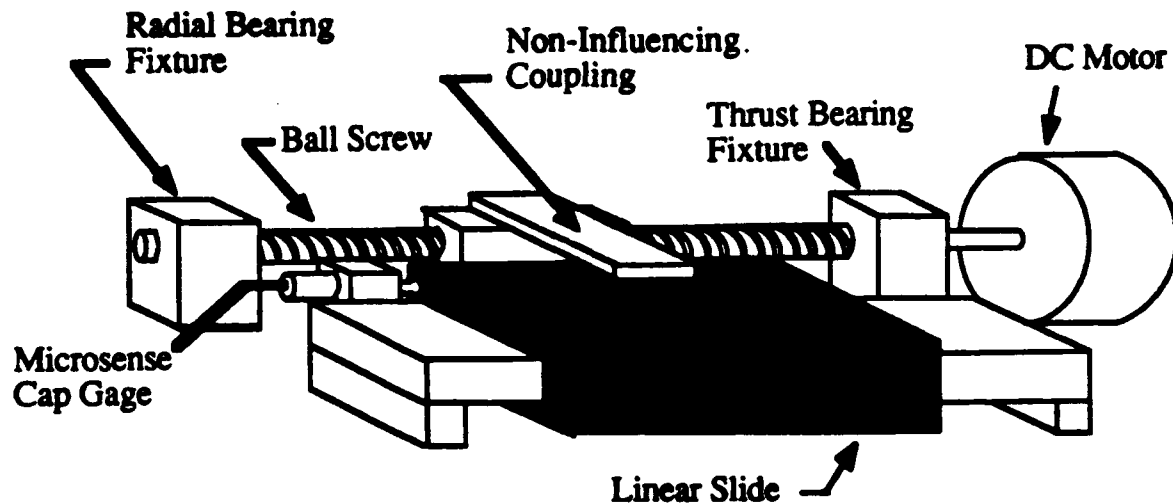


Figure 1: Experimental apparatus

14.2 CHARACTERIZATION

The experimental apparatus illustrated in Figure 1 consists of a DC servo motor, precision ball screw, and air-bearing slideway. Although the offset ball-screw configuration is not optimal, the apparatus is suitable as a testbed for investigating ball-screw dynamics at small displacements. A linear amplifier and analog current regulator are used to drive the motor. Position feedback is through a capacitance sensor which provides an analog signal to the controller. The mechanical components are mounted on an air-table for vibration isolation and are located in a temperature controlled room. All of the components, with exception of the current regulator and bearing supports, are commercially available.

Open loop testing indicates that small angle ball screw rotation (and subsequent slide displacement) occurs for any applied torque. The displacement vs. torque relationship illustrated in Figure 2 was obtained by slowly increasing the applied torque. The relationship in Figure 2 is surprising because it contradicts the initial intuitive notion, based on classical stick/slip friction models; that is, no rectilinear ball-nut displacement will occur until the applied torque exceeds some threshold. Figure 2 suggests that there is a region of sub-micrometer motion resulting from elastic deformation. Such "spring" behavior appears to be a consequence of elastic deformation of the contact patches on the balls, according to a geometric model of the ball-nut developed by Cuttino and described in Section 5 and reference [4]. However, as the applied torque increases, a friction breakaway is observed and the deformation remains constant.

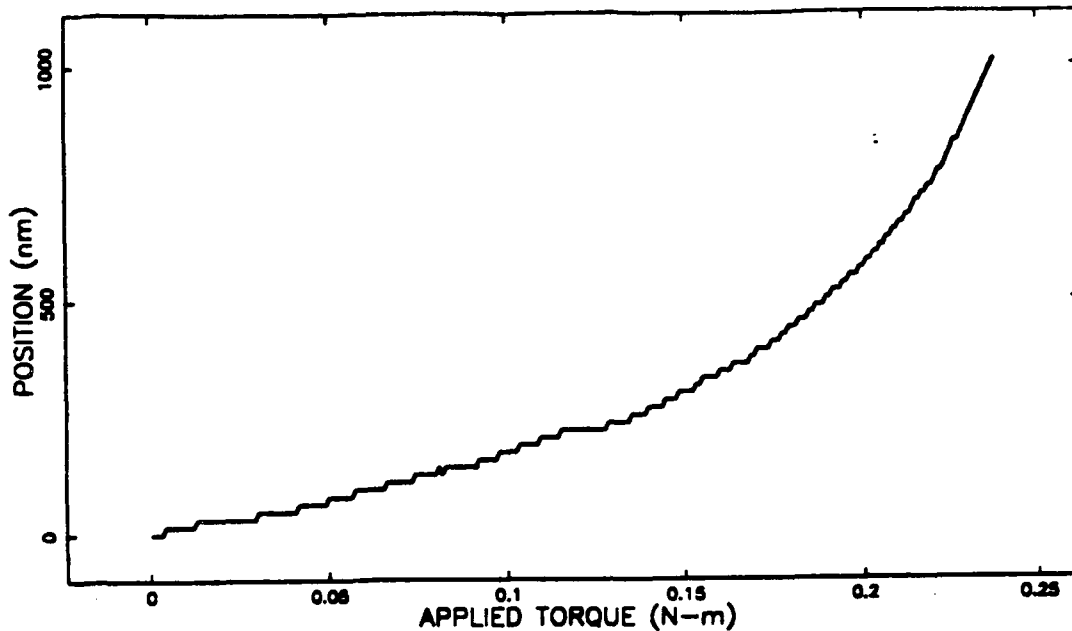


Figure 2: Displacement vs. applied torque relationship of ball-screw.

If such elastic deformation does occur, a corresponding resonant peak should appear in the open-loop frequency response. In fact, by linearly approximating the nonlinear displacement-torque relationship for sub-micrometer motion, the resonant frequency should decrease as the range of motion increases. This happens because the linear approximation results in a decreasing spring stiffness as the slide displacement increases. The frequency response measurements in Figures 3 and 4 confirm this prediction. While the dominant 125 Hz peak resulting from the coupling between the ball-nut and slide remains stationary, the microdynamic peak moves from 180 Hz down to 143 Hz as the slide displacement increases from approximately 25 nm to 1500 nm.

These peculiar dynamics can also be illustrated via the position step response shown in Figures 5 and 6. In this case a PI position controller was applied and the gains were deliberately set to create a slow response to the 20 μm step. Figure 5 suggests typical stick/slip behavior, yet the magnified view of the initial response in Figure 6 shows nanometer motion prior to the friction breakaway while the integral controller term is increasing the motor torque.

The complexities of the observed dynamics force reevaluation of the accepted macroscopic ball-screw models which include classical stick/slip friction.

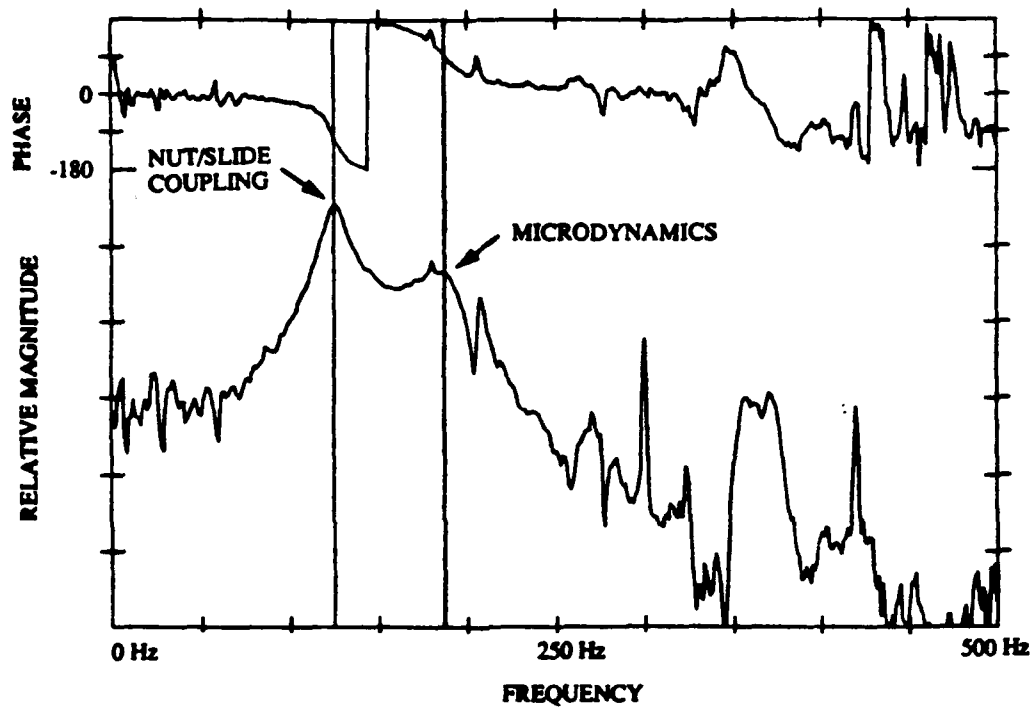


Figure 3: Small amplitude microdynamic transfer function of ball-screw driven slide. The maximum slide displacement is approximately 25 nm.

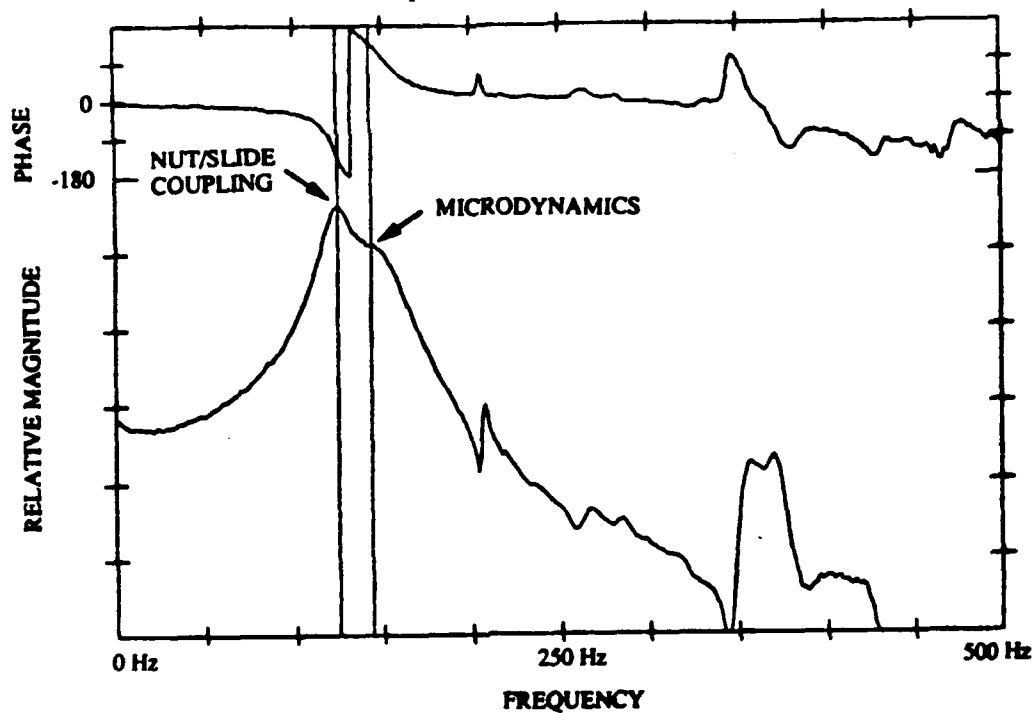


Figure 4: Large amplitude microdynamic transfer function of ball-screw driven slide. The maximum slide displacement is approximately 1500 nm.

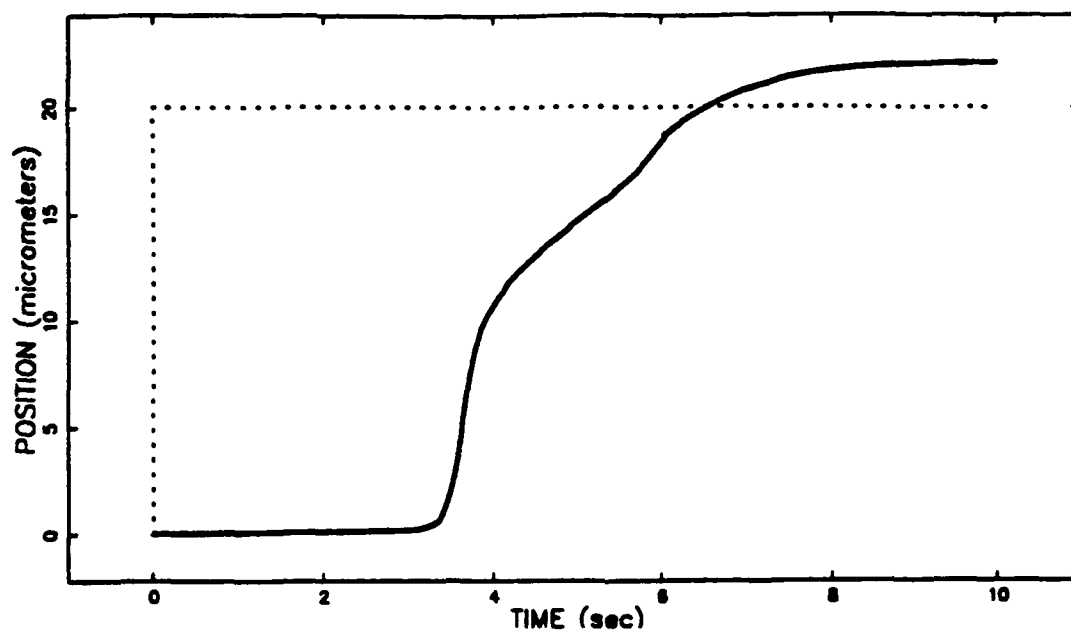


Figure 5: 20 μm step response of apparatus while under PI position control.

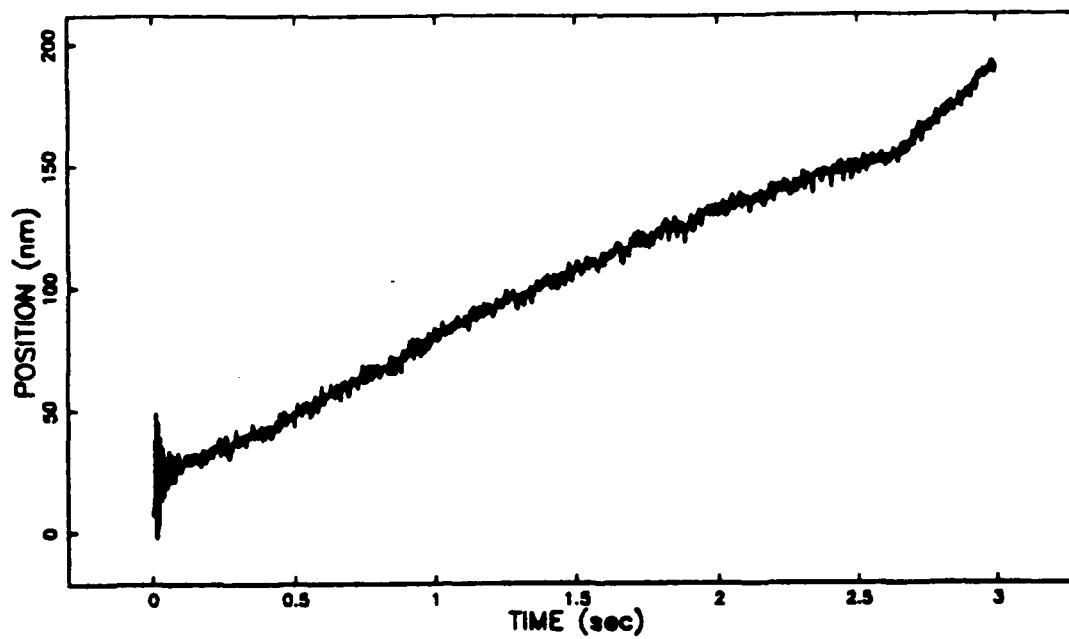


Figure 6: Magnified view of the initial region of the step response shown in Figure 5.

14.3 MODELING

The simplest conceptual model that can explain most of the observed behavior is the goal of this study. The model developed attaches a classical stick/slip friction interface to a globally linear ball-screw model via a parallel combination of a spring and damper. A conceptual diagram of this model is included in Figure 7. This diagram is not intended to represent the actual mechanical components of the system, rather the ideal frictionless ball-nut, spring, damper, and friction interface collectively model the actual ball-nut. Obviously the model includes some simplifying assumptions, such as a rigid ball-screw, a rigid coupling between ball-nut and slide, and a frictionless air-bearing slideway. However, these assumptions are made because the goal is to concentrate on the dynamics of the ball-nut induced by friction.

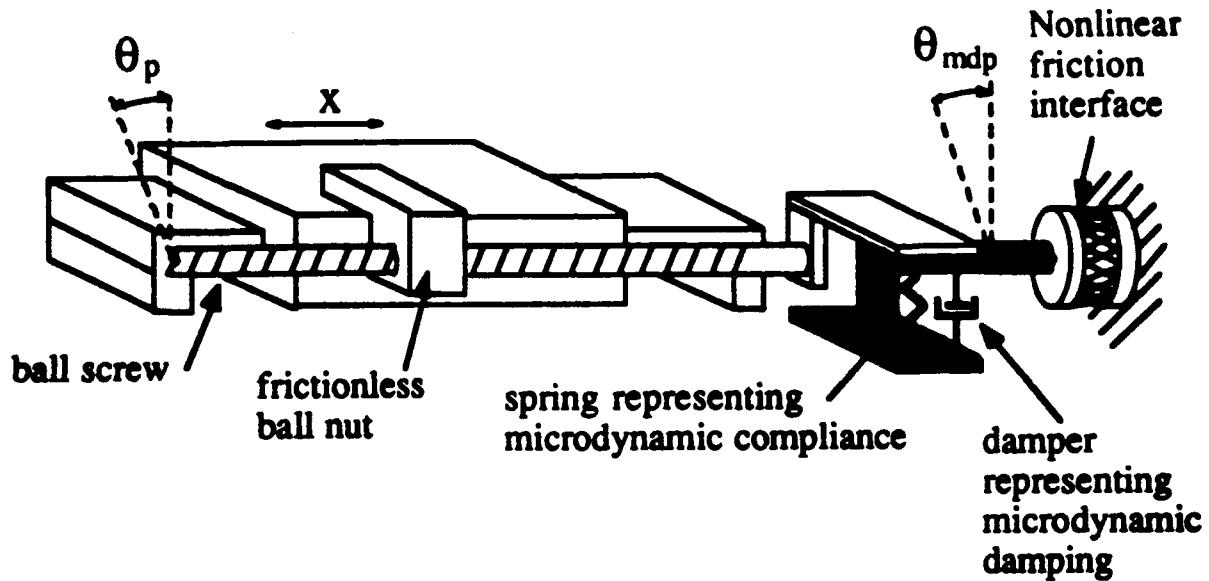


Figure 7: Conceptual diagram illustrating the structure of the model.

The comprehensive dynamics of the model in Figure 7 are described by

$$J\ddot{\theta}_p + B\dot{\theta}_p + B_{fr}(\dot{\theta}_p - \dot{\theta}_{mdp}) + K_{fr}(\theta_p - \theta_{mdp}) = u_l + u_a \quad (1)$$

Where J is the rotational inertia of the motor armature and ball screw, B is the damping coefficient of the motor and ball screw with respect to ground, B_{fr} is the damping coefficient of the ball screw with respect to the microdynamic reference position θ_{mdp} , K_{fr} is the microdynamic spring constant, θ_p is the position of the ball screw, and $u_l + u_a$ is the torque generated by the motor. The displacement of the slide is

$$x = K_p \theta_p \quad (2)$$

where the lead of the screw is K_p .

The specific dynamic modes are determined according to

$$\dot{\theta}_{\text{mdp}} = \begin{cases} 0 & \text{(microdynamics)} \\ \dot{\theta}_p & \text{(macrodynamics)} \end{cases} \quad (3)$$

When the plant is in the microdynamic mode the friction interface is stuck while the macrodynamic mode occurs when there is sliding across the interface and the deformation of the microdynamic spring is constant. In the macrodynamic case, the friction torque is

$$\tau_f = K_{fr}(\theta_p - \theta_{\text{mdp}}) \quad (4)$$

The plant will always be in the microdynamic mode when motion is initiated and will switch into the macrodynamic mode if the torque applied to the friction interface exceeds some threshold. Once switched into the macrodynamic mode, the plant will only revert to the microdynamic mode if the ball screw reverses direction. Such an action causes the microdynamic spring to relax and the torque transmitted to the friction interface to drop below the stiction threshold.

The two-stage dynamics observed in Figures 5 and 6 and the two-stage plant model suggest that the controller will also have multiple stages.

14.4 CONTROLLER DESIGN OBJECTIVE

Given the magnitude dependent dynamics exhibited in Figures 3 and 4, the objective of the controller design is to provide a predictable linear closed loop response for nanometer slide displacements. The MRAC technique includes a reference model which generates an ideal response and adaptive laws which adjust variable parameters to ensure that the plant output follows the reference model output, regardless of unknown plant parameters [7, 8, 11]. Restrictions do apply to the structure of the reference model and lack of feedback information on the plant states can hinder implementation of the adaptive laws. However, applications have shown the effectiveness of the technique [10, 12].

14.4.1 Design of Microdynamic Adaptive Controller

A PI control law was implemented to determine the effect the microdynamic nonlinearity has on the closed-loop response. The gains were adjusted experimentally to provide a rise time of approximately 0.1 second and a slightly overdamped response to a 300 nm step command. However, when the same controller gains are used for different step magnitudes, the closed-loop response varies significantly. Figures 8(a) and 8(b) show the closed-loop response varying from over-damped to under-damped as the step magnitude increases.

The problem of controller design illustrated in Figures 8(a) and 8(b) is understood by noting that the microdynamic plant model includes two parameters that are critical to controller design but are not well known and may be time-varying. These are the microdynamic reference position (θ_{mdp}) and the microdynamic spring constant (K_{fip}). The uncertainty of K_{fip} results from linearizing the nonlinear displacement-torque relationship in Figure 1. Likewise, the reference position (θ_{mdp}) is also difficult to measure.

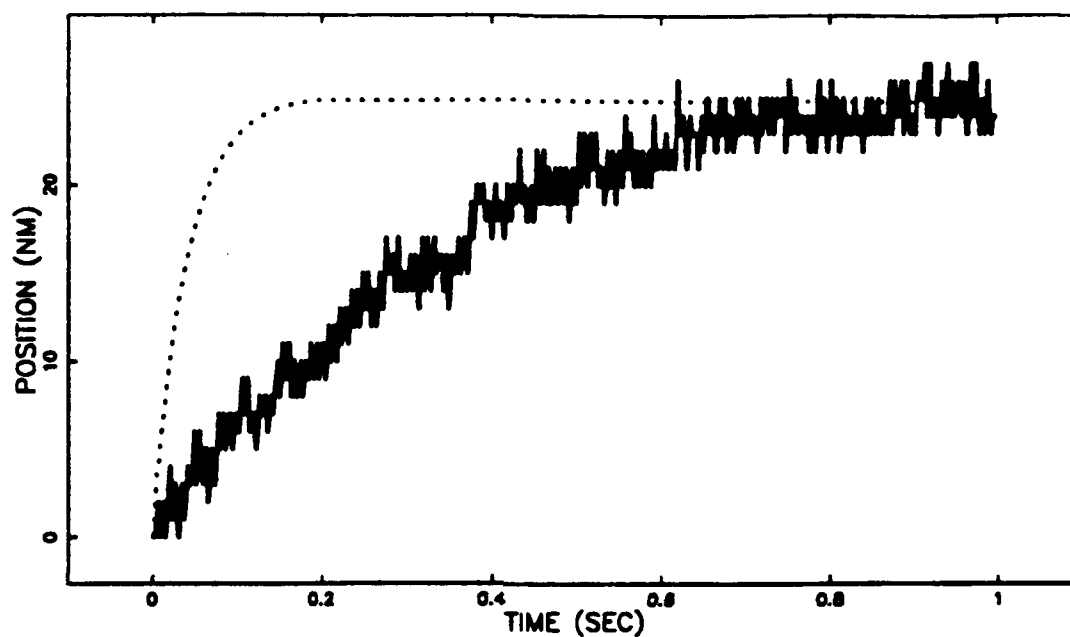


Figure 8(a): Response of plant to 25 nm step command while under PI control.

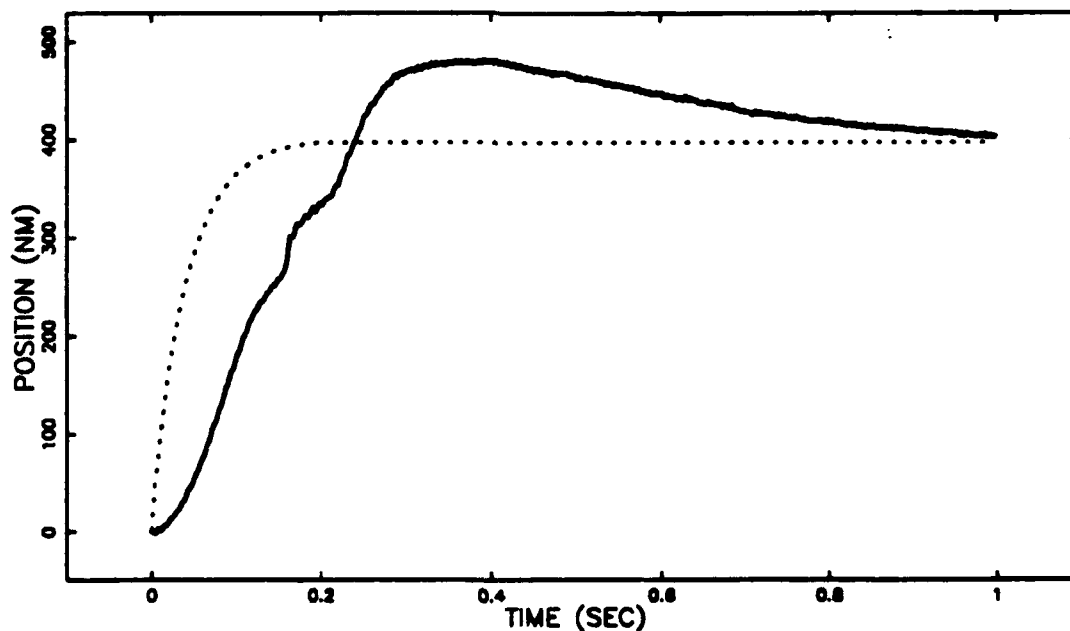


Figure 8(b): Response of plant to 400 nm commanded step while under PI control.

The model's microdynamic spring constant (K_{frm}) was obtained by linearizing the torque-displacement relationship and the initial estimate of B_{frm} was obtained by matching experimental closed-loop response data to computer simulation output. The reference position (θ_{mdm}) can be approximated by the ball screw position (θ_p) when the applied torque (u_1+u_2) is zero.

The dynamics of the amplifier, armature circuit, and current regulator are neglected, as they are much faster than the mechanical dynamics of the plant. This assumption allows the output of the linear controller (u_1) and the output of the adaptive scheme (u_2) to be modeled as torque inputs to the reference model and plant.

The reference model is

$$\dot{\underline{\theta}}_m = \underline{A}_m \underline{\theta}_m + \underline{B}_m \underline{u}, \quad (5)$$

where

$$\underline{\theta}_m = \begin{bmatrix} \theta_{1m} \\ \theta_{2m} \end{bmatrix}, \quad (6)$$

$$\underline{A}_m = \begin{bmatrix} 0 & 1 \\ -\frac{K_{frm}}{J} & -\frac{(B+B_{fr})}{J} \end{bmatrix}, \quad (7)$$

$$\underline{B}_m = \begin{bmatrix} 0 & 0 \\ \frac{1}{J} & \frac{K_{frm}\theta_{mdm}}{J} \end{bmatrix}, \quad (8)$$

and

$$\underline{u} = \begin{bmatrix} u_1 \\ 1 \end{bmatrix}. \quad (9)$$

The MRAC scheme is designed to force the plant to follow the reference model despite poor knowledge of the parameters described above. Figure 9 is a block diagram of the structure of a typical MRAC controller, where the linear controller could be a state variable controller as all states of the reference model are observable.

The input to the reference model, u_1 , is generated by the linear microdynamic controller. In this case, a PI law controls the reference model.

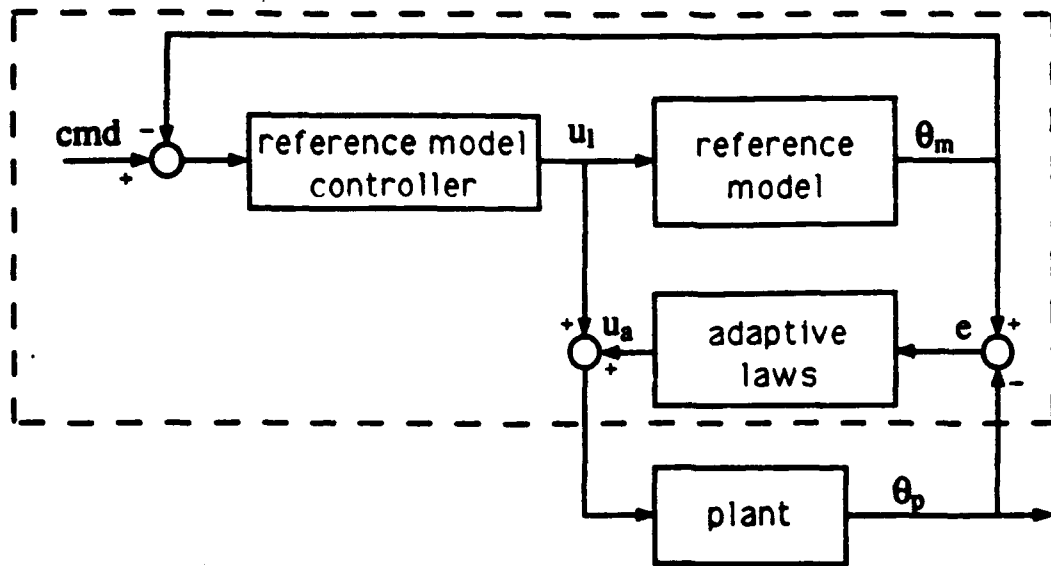


Figure 9: Block diagram of an MRAC scheme.

The unknown plant parameters can be expressed in terms of the corresponding reference model parameters and unknown differences:

$$K_{frp} = K_{frm} + \Delta K_{fr}, \quad (10)$$

$$\theta_{mdp} \approx \theta_{mdm} + \Delta\theta_{md}. \quad (11)$$

The adaptive input is a function of the variable parameters, plant states, and plant inputs:

$$u_a = P_1 \theta_p + P_2. \quad (12)$$

Using Equations (1), (3), (10), (11), and (12), the microdynamic plant is

$$\dot{\theta}_p = A_p \theta_p + B_p u, \quad (13)$$

where

$$\theta_p = \begin{bmatrix} \theta_{1p} \\ \theta_{2p} \end{bmatrix}, \quad (14)$$

$$\Delta_p = \begin{bmatrix} 0 & 1 \\ -\frac{(K_{frm} + \Delta K_{fr} - P_1)}{J} & -\frac{(B+B_{fr})}{J} \end{bmatrix}, \quad (15)$$

$$B_p = \begin{bmatrix} 0 & 0 \\ \frac{1}{J} \frac{(K_{frm} + \Delta K_{fr})(\theta_{mdm} + \Delta \theta_{md}) + P_2}{J} \end{bmatrix}, \quad (16)$$

and

$$\underline{u} = \begin{bmatrix} u_1 \\ 1 \end{bmatrix}. \quad (17)$$

The error dynamics are obtained by subtracting Equation (13) from Equation (5).

$$\dot{\underline{\epsilon}} = \underline{A}_m \underline{\epsilon} + (\underline{A}_m - \underline{A}_p) \underline{\theta}_p + (\underline{B}_m - \underline{B}_p) \underline{u}, \quad (18)$$

$$\underline{\epsilon} = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \underline{\theta}_m - \underline{\theta}_p \quad (19)$$

$$\underline{A} = \underline{A}_m - \underline{A}_p \quad (20)$$

$$\underline{B} = \underline{B}_m - \underline{B}_p \quad (21)$$

Application of the design process outlined by Gilbert, et. al. [8], results in the Lyapunov function

$$\begin{aligned} V = \underline{\epsilon}^T \underline{Q} \underline{\epsilon} + \frac{1}{\alpha_{21}} [a_{21} + \beta_{21}(e_1 q_{12} a_{21} \theta_{1p} + e_2 q_{22} a_{21} \theta_{1p})]^2 \\ + \frac{1}{\gamma_{22}} [b_{22} + \delta_{22}(e_1 q_{12} b_{22} u_2 + e_2 q_{22} b_{22} u_2)]^2 \end{aligned} \quad (22)$$

The symmetric matrix \underline{Q} is determined by the equation

$$\underline{A}_m^T \underline{Q} + \underline{Q} \underline{A}_m = -\underline{P} \quad (23)$$

where \underline{P} is arbitrary positive definite symmetric. The a_{ij} and b_{ij} are elements of \underline{A} and \underline{B} , respectively. The parameters α_{ij} and γ_{ij} are arbitrary positive definite constants while the β_{ij} and δ_{ij} are arbitrary positive semi-definite constants.

The general form of the adaptive laws are then chosen to force the time derivative of Equation (22) to be negative definite over the error space $[e_1 \ e_2]^T$. This insures that the model following error goes to zero as time goes to infinity. For this particular application, the adaptive laws are

$$P_1(t) = \alpha_{21} \int_0^t (eq_{12}\theta_p + \dot{eq}_{22}\theta_p) dt + \beta_{21} (eq_{12}\theta_p + \dot{eq}_{22}\theta_p) + P_1(t_0) \quad (24)$$

$$P_2(t) = \gamma_{22} \int_0^t (eq_{12} + \dot{eq}_{22}) dt + \delta_{22} (eq_{12} + \dot{eq}_{22}) + P_2(t_0) \quad (25)$$

The inclusion of the model following position error (e) in Equations (24) and (25) is a result of the particular Lyapunov function presented by Gilbert, et.al. Other Lyapunov design techniques do not necessarily result in adaptive laws which make use of the position error [10]. It is critical in this application that the position signal be fully utilized, as the velocity is so low that the tachometer is of little use. A psuedo-derivative is calculated from the position signal, but this may be sensitive to sensor noise.

The "proportional plus integral" form of the adaptive laws allows the controller designer to influence the manner in which the plant converges to the model by varying the values of the arbitrary positive constants ($\alpha_{21}, \beta_{21}, \gamma_{22}, \delta_{22}$). The integral terms ultimately force the error to zero, but the proportional terms speed up the adaptation process. Again, the presence of the proportional and integral terms results from the specific form of the Lyapunov.

14.4.2 Design of Macrodynamic Controller

Because the macrodynamic kinetic friction in the plant can result in a nonlinear closed-loop response, the goal is an adaptive controller which generates an adaptive feedforward term intended to cancel the friction torque described by

$$\tau_f = \tau_{kf} \operatorname{sgn}(\dot{\theta}_p) \quad (26)$$

This approach has already been studied, but is included in this investigation to offer a comprehensive control approach for the nonlinear plant dynamics [10, 12]. The applicable plant equation

$$(J_m + J_s) \ddot{\theta}_p + B_r \dot{\theta}_p + \tau_{kf} \operatorname{sgn}(\dot{\theta}_p) = u_l + u_a \quad (27)$$

is obtained by setting

$$\dot{\theta}_{mdp} = \dot{\theta}_p \quad (28)$$

in Equation (1). The structure of this MRAC scheme is the same as that shown in Figure 9, and the adaptive input is expressed by

$$u_a = P_3 \operatorname{sgn}(\dot{\theta}_p), \quad (29)$$

where P_3 is the variable parameter which corresponds to the unknown kinetic friction torque (τ_{kf}).

After substituting in Equation (29), Equation (27) can be written in standard form as

$$\dot{\underline{\theta}}_p = \underline{A}_p \underline{\theta}_p + \underline{B}_p \underline{u}, \quad (30)$$

where

$$\underline{\theta}_p = \begin{bmatrix} \theta_{1p} \\ \theta_{2p} \end{bmatrix}, \quad (31)$$

$$\underline{A}_p = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{B_r}{(J_m + J_s)} \end{bmatrix}, \quad (32)$$

$$\underline{B}_p = \begin{bmatrix} 0 & 0 \\ \frac{1}{(J_m + J_s)} & \frac{P_3 - \tau_{kf}}{(J_m + J_s)} \end{bmatrix}, \quad (33)$$

$$\underline{u} = \begin{bmatrix} u_1 \\ \operatorname{sgn}(\dot{\theta}_p) \end{bmatrix}. \quad (34)$$

The reference model is the same structure as the plant model, but has no nonlinear friction and can be expressed as

$$\dot{\underline{\theta}}_m = \underline{A}_m \underline{\theta}_m + \underline{B}_m \underline{u}, \quad (35)$$

where

$$\underline{\theta}_m = \begin{bmatrix} \theta_m \\ \dot{\theta}_m \end{bmatrix} = \begin{bmatrix} \theta_{1m} \\ \theta_{2m} \end{bmatrix}, \quad (36)$$

$$\Delta_m = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{B_r}{(J_m + J_s)} \end{bmatrix}, \quad (37)$$

$$B_m = \begin{bmatrix} 0 & 0 \\ \frac{1}{(J_m + J_s)} & 0 \end{bmatrix}, \quad (38)$$

$$u = \begin{bmatrix} u_1 \\ \text{sgn}(\dot{\theta}_p) \end{bmatrix}. \quad (39)$$

Defining the model following error as

$$e = \theta_m - \theta_p = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}, \quad (40)$$

and applying Gilbert's MRAC design technique results in the specific Lyapunov function

$$V = e^T Q e + \frac{1}{\gamma_{22}} [b_{22} + \delta_{22}(e_1 q_{12} b_{22} u_2 + e_2 q_{22} b_{22} u_2)]^2, \quad (41)$$

and the adaptive law

$$P_3 = K_{22}^b = \gamma_{22} \int_0^t (e_1 q_{12} u_2 + e_2 q_{22} u_2) dt + \delta_{22} (e_1 q_{12} u_2 + e_2 q_{22} u_2) + K_{22}^b(t_0) \quad (42)$$

The Q matrix can be found by solving Equation (23) while γ_{22} , and δ_{22} are arbitrary positive adaptive gains.

14.5 DSP IMPLEMENTATION

The adaptive control equations, linear microdynamic controller equations, and reference model equations were converted from continuous-time to discrete-time via the bilinear approximation. Computer simulations written in ACSL (Advanced Continuous Simulation Language) were used to check the resulting difference equations. The control scheme was implemented on a hierarchical architecture which uses a Texas Instruments TMS320C30 digital signal processor [13]. The control code is written in C and installed as a routine which services a 2 kHz timer driven interrupt.

14.6 EXPERIMENTAL RESULTS

Given that the derived controller has two stages, the goal is to initiate *in situ* switching between the stages. However, as a first attempt, the controller mode was set *a priori* according to the magnitude of the command.

Figures 10 and 11 show the response of the plant to 25 nm and 400 nm steps while under control of the microdynamic MRAC scheme. In this case the linear microdynamic controller was tuned to give the reference model a slightly overdamped response with a rise time of approximately 0.1 seconds. Note that Figures 10 and 11 display very similar responses, regardless of the magnitude of the step while the closed-loop responses in Figures 8(a) and 8(b) obtained by PI control show considerably different responses.

Although Figures 10 and 11 suggest that the microdynamic controller stage is effective for sub-micrometer motion, the experiments indicated the performance of this controller deteriorated dramatically as the range of motion increased beyond 400 nm. Figures 12 and 13 include a 500 nm step responses while under microdynamic and macrodynamic MRAC control, respectively. Note that neither of the controllers perform well in this range, although Figures 14 and 15 indicate that the macrodynamic controller performs well for steps greater than 1 μm .

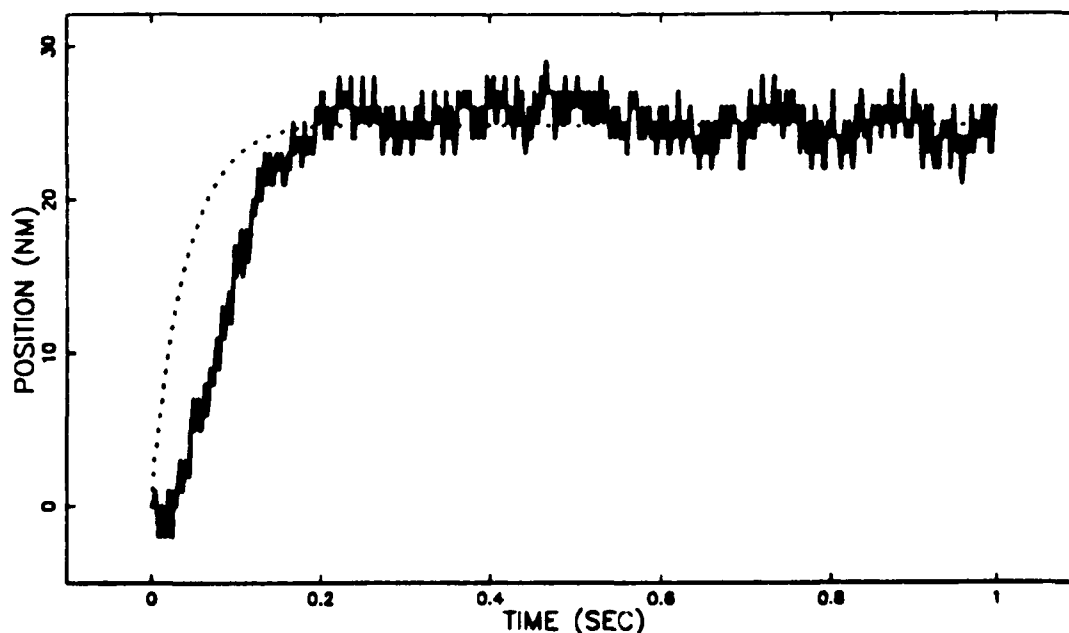


Figure 10: Response of plant to 25 nm step command while under microdynamic adaptive control. The dotted line is the position output of the reference model.

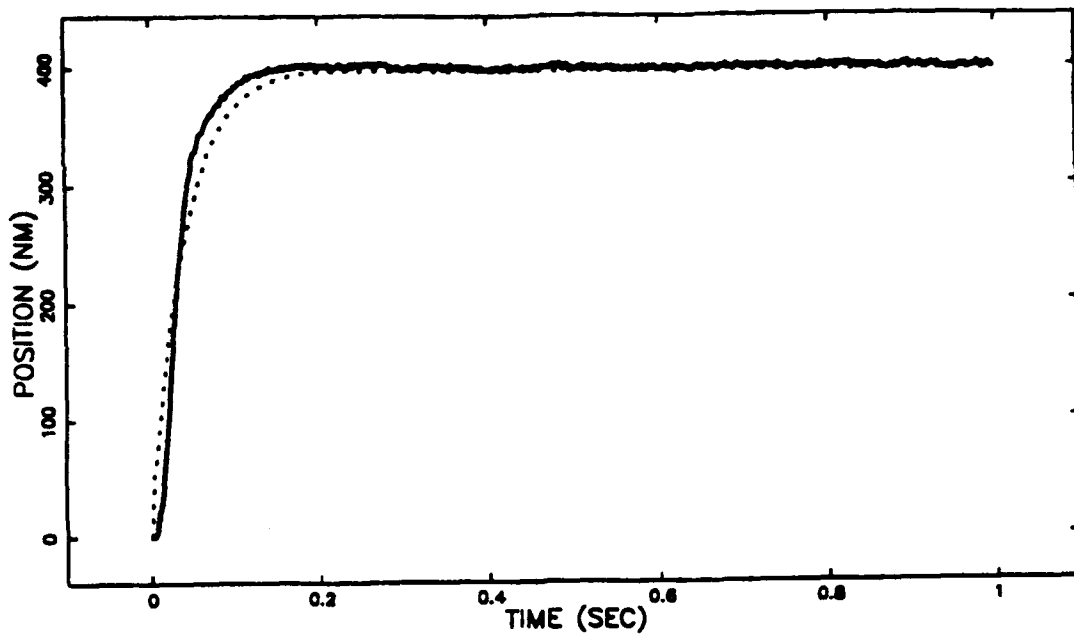


Figure 11: Response of plant to 400nm step command while under microdynamic adaptive control. The dotted line is the position output of the reference model.

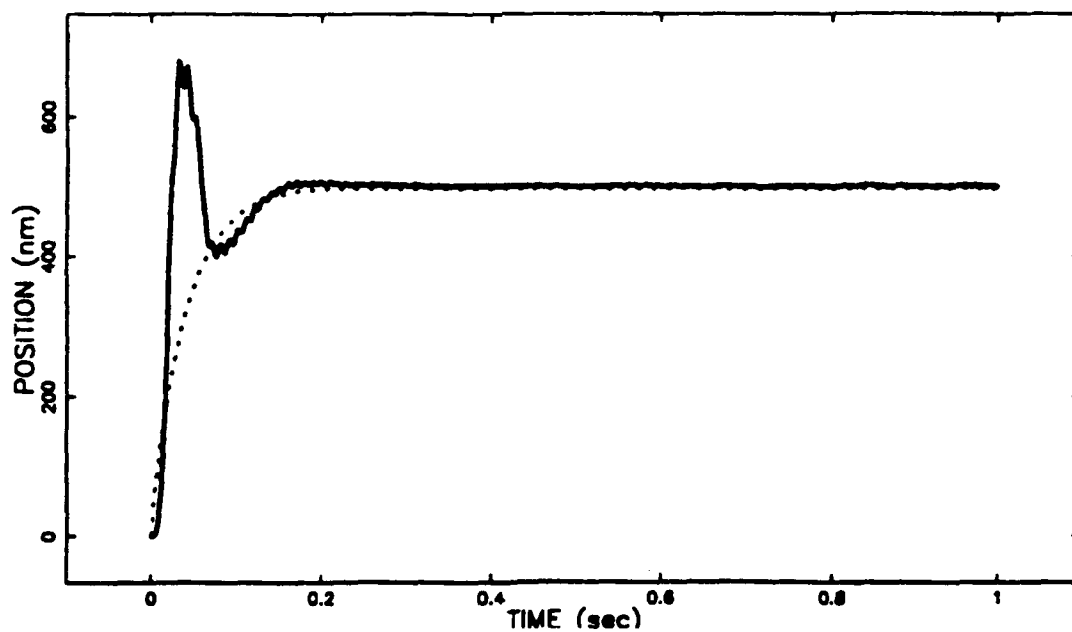


Figure 12: Response of plant to 500 nm step command while under microdynamic adaptive control. The dotted line is the position output of the reference model.

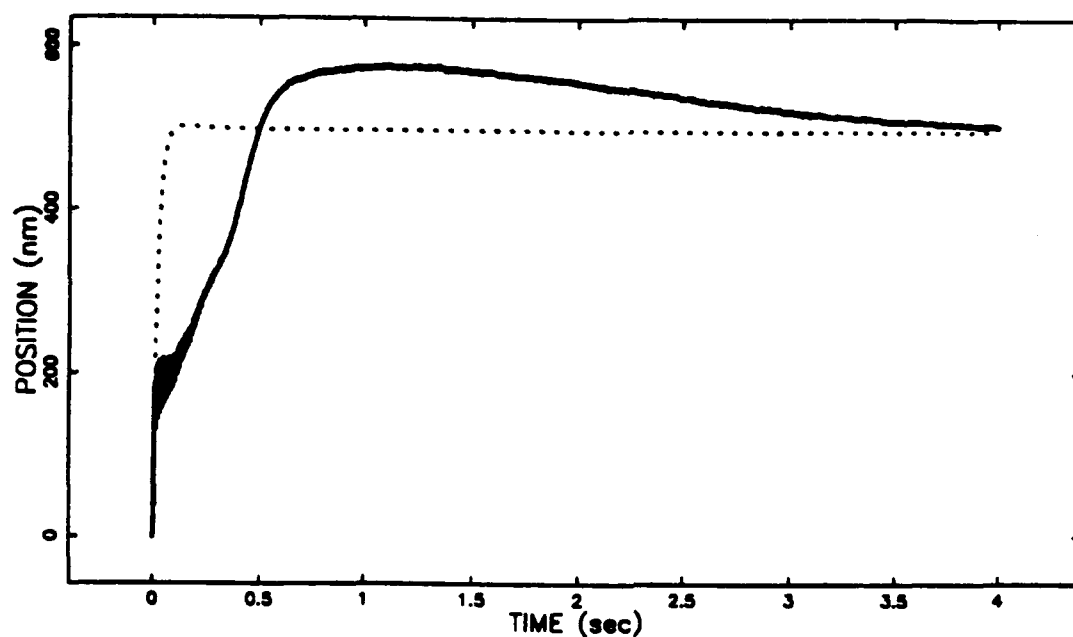


Figure 13: Response of plant to 500 nm step command while under macrodynamic adaptive control. The dotted line is the position output of the reference model.

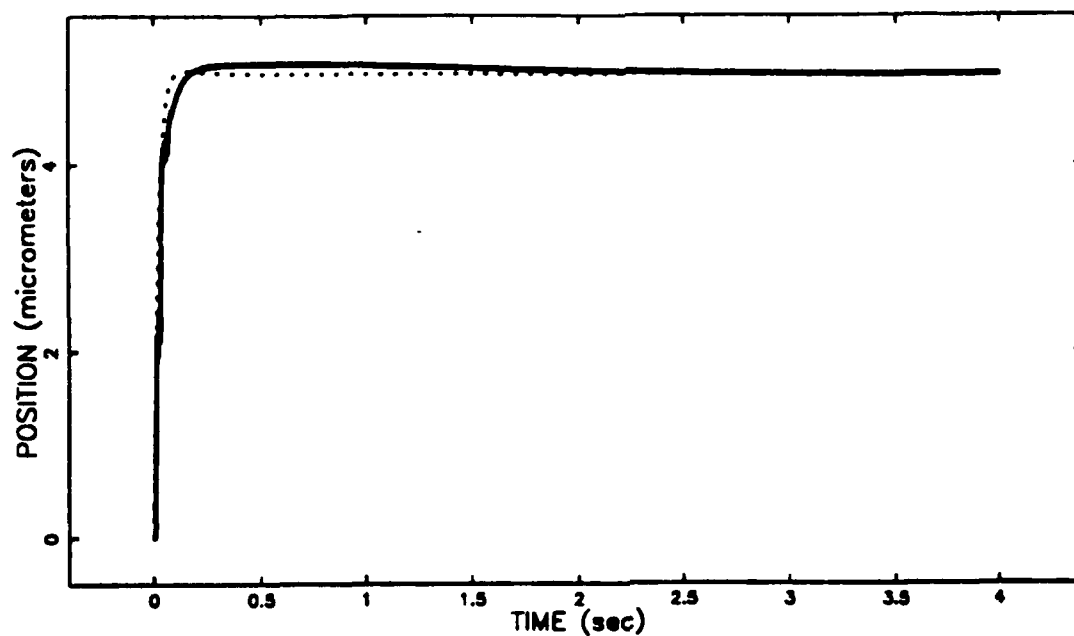


Figure 14: Response of plant to 5 μm step command while under macrodynamic adaptive control. The dotted line is the position output of the reference model.

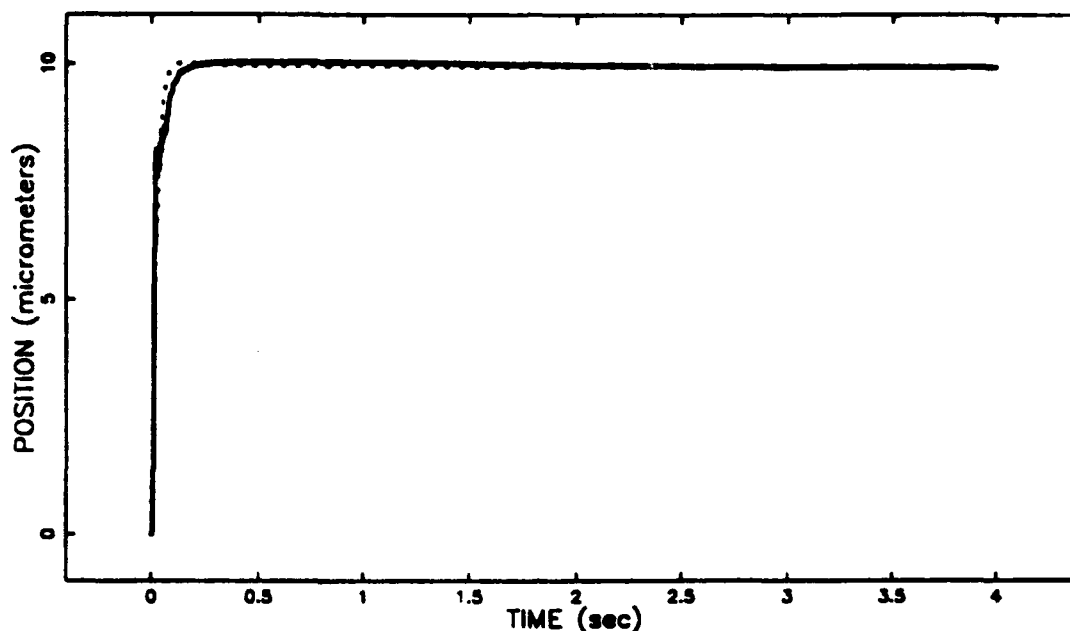


Figure 15: Response of plant to 10 μm step command while under macrodynamic adaptive control. The dotted line is the position output of the reference model.

The presence of a range (500 nm to 1 μm) in which neither controller stage performs well can be understood by realizing that the multi-stage controller was designed around a piecewise linear plant model which becomes increasingly inaccurate near the boundary between the linear segments. If either controller stage is applied while the plant is operating near this boundary, then the controller will be attempting to control a plant which does not correspond well with the model for which it was designed.

14.7 CONCLUSION

The ball-screw undoubtedly exhibits nonlinear (magnitude dependent) dynamics which can have a profound effect on the performance of the drive using classical control schemes. The sub-micrometer dynamics (microdynamics) are characterized by a friction induced nonlinear elastic behavior which allows nanometer motion without incurring a friction breakaway. The larger scale dynamics (macrodynamics) are characterized by a classical kinetic (sliding) friction model.

The model reference adaptive control approach has been applied to the microdynamic (pre friction breakaway) and macrodynamic modes of the dual-mode dynamics observed experimentally. Each controller stage performs well within its respective range of motion, but neither performs well when operated near the transition between microdynamic and macrodynamic plant modes.

Although classical control laws, such as PI, provide acceptable performance in many applications, the observed nonlinear dynamics will limit their potential when used for nanometer motion control. Within the microdynamic region, the adaptive control is more robust with respect to the nonlinear dynamics than a PI controller.

Future work will include incorporating the form of the plant nonlinearity directly into the controller design to create a single stage, nonlinear controller that can operate the plant to nanometer accuracies over a range of centimeters without changing gains or structure.

References

1. DeWeerth, S.P., L. Nielsen, C.A. Mead, and K.J. Astrom, "A Simple Neuron Servo", *IEEE Transactions on Automatic Control*, 2(2), p. 248-251, 1991.
2. Southward, S.C., C.J. Radcliffe and C.R. MacCluer, "Robust Nonlinear Stick-Slip Friction Compensation", submitted to *Journal Dynamic Systems, Measurement, and Control*, , 1990.
3. Yang, S., and M. Tomizuka, "Adaptive Pulse Width Control for Precise Positioning Under the Influence of Stiction and Coulomb Friction", *Journal of Dynamic Systems, Measurement, and Control*, 110, p. 221-227, 1988.
4. Cuttino, J., "The ball screw as an actuator for nanometer motion", submitted to *Precision Engineering*, , 1991.
5. Courtney-Pratt, J., and E. Eisner, "The effect of a tangential force on the contact of metallic bodies", *Proceedings of the Royal Society, A*, 238: p. 529-550, 1957.
6. Futami, S., A. Furutani and S. Yoshida, "Nanometer Positioning and its Microdynamics", *Nanotechnology*, 1, p. 31-37, 1990.
7. Landau, Y.D., "Adaptive control, the model reference approach", *New York: Marcel Dekker, Inc.*, 1979.
8. Landau, I.D., "A Survey of Model Reference Adaptive Techniques - Theory and Applications", *Automatica*, 10, p. 353-379, 1974.
9. Gilbert, J.W., R.V. Monopoli and C.F. Price, "Improved Convergence and Increased Flexibility in the Design of Model Reference Adaptive Control Systems" in *Proceedings of Ninth Symposium on Adaptive Processes*, 1970. IEEE.
10. Gilbert, J.W., and G.C. Winston, "Adaptive Compensation for an Optical Tracking Telescope", *Automatica*, 10, p. 125-131, 1974.
11. Lindorff, D.P., and R.L. Carroll, "Survey of adaptive control using Liapunov design", *International Journal of Control*, 18(5): p. 897-914, 1973.
12. Canudas, C., K.J. Astrom and K. Braun, "Adaptive Friction Compensation in DC Motor Drives" in *Proceedings of IEEE International Conference on Robotics and Automation*, 1986. San Francisco, CA: IEEE.
13. Fornaro, R.J., K.P. Garrard, and L.W. Taylor, "Architectures and Algorithms for Computer Control of High Precision Machine Tools" in *Proceedings of ASPE Annual Conference*, 1990. Rochester, New York.

15 CONTROL OF PRECISION SLIDE MOTION FOR VIBRATION REDUCTION

Jeffrey A. Abler

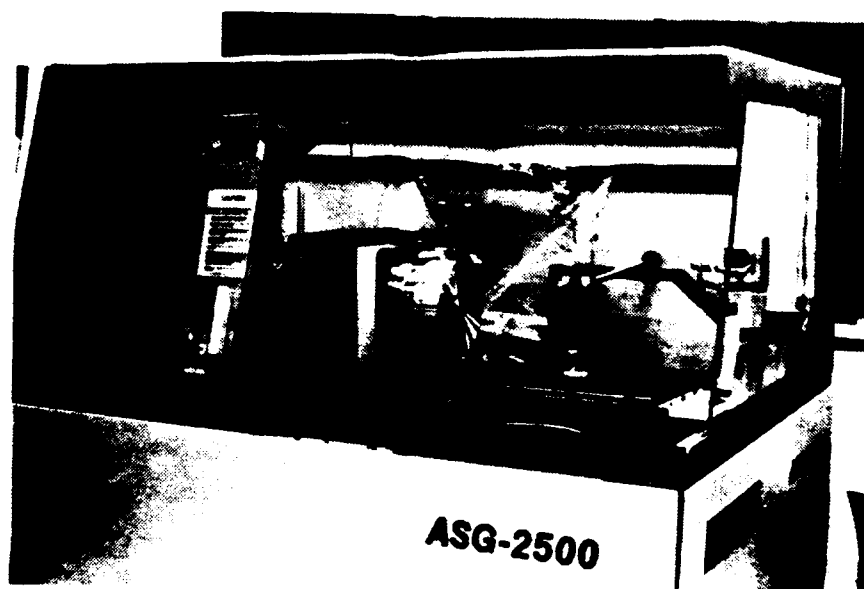
Graduate Research Assistant

Paul I. Ro

Assistant Professor

Department of Mechanical and Aerospace Engineering

The highly underdamped nature of a low friction hydrostatic slide makes the system to which it belongs susceptible to vibrations caused by disturbances acting on the system. When driving a lightly damped system with a servo motor, the design of a controller which rejects the vibration producing disturbances can be undermined by the need to maintain the system bandwidth to achieve good tracking performance. The use of classical lag/lead compensation and conventional filtering techniques to dampen the oscillatory mode often has adverse effects on the system bandwidth. A control design procedure is developed and implemented on the diamond turning machine which separates the design of the system bandwidth from that of the damping of the highly underdamped mode. Using the scheme, the slide vibration is significantly less than when using conventional PID control, with a corresponding improvement in the surface roughness of machined parts.



15.1 INTRODUCTION

The accuracy of a diamond turned part is greatly dependent on the ability to control the relative position of the tool and the workpiece. Precision slide systems for diamond turning are often designed with a hydrostatic oil or air bearing to reduce friction. Because of the small amount of friction, the resulting system is highly underdamped in the direction of motion. Such a system is susceptible to vibration caused by the presence of external disturbances such as structural vibration, acoustic noise, and electrical noise. The slide vibration is directly transferred to the machined parts producing increased surface roughness. Similarly, errors in tracking performance (deviations of slide position from the intended path), can result in figure errors in the parts machined. Given these two sources of error in diamond turning, it is desirable to control the slides utilizing a scheme which dampens the slide vibration while providing good tracking performance.

Previously, various conventional schemes [1,2] were analyzed and tested as position loop controllers on a Pnuemo ASG-2500 diamond turning machine. PID control, lag/lead compensation, notch and Butterworth filters were tested over the range of practical machining feedrates. All of these schemes were unable to simultaneously achieve both of the desired goals. In each case the problem stemmed from the interaction of the underdamped slide dynamics with the dynamics of the motor and amplifier which drive the slide. The incorporation of a control scheme in the position loop will effect both sets of dynamics. The general problem encountered involves the competing factors of the amount of lead/lag added to the system to improve the dynamics of the motor and amplifier (improving system bandwidth and time response), versus the amount of lag/lead added to the system to dampen the lightly damped poles of the hydrostatic slide (reducing vibration). In each case, the improvement of one had a diminishing (or adverse) effect on the other.

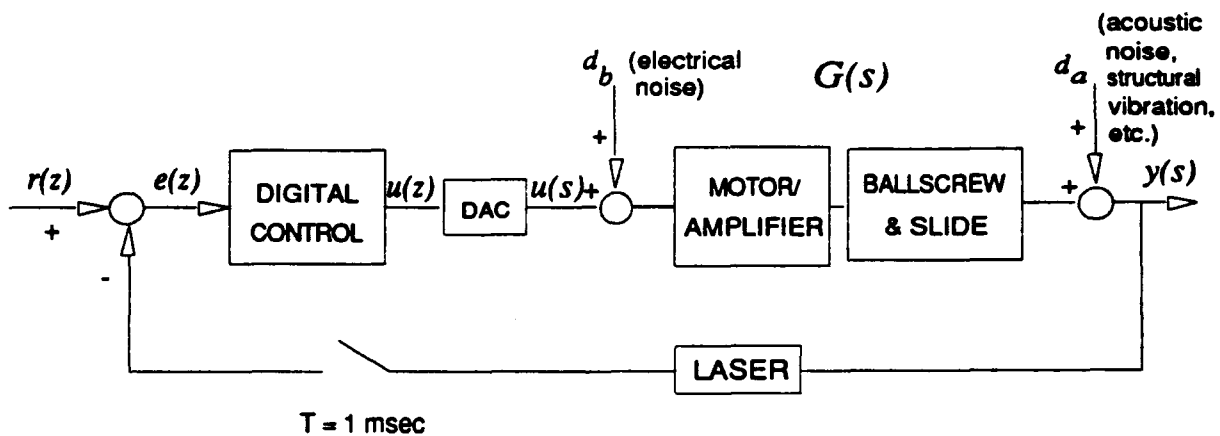
A control scheme was developed (Zero Placement Integral Control) which balanced the improvement of the system bandwidth with the damping of lightly damped slide [1]. However, the procedure used to design the scheme was somewhat *ad hoc*, and a more mathematically explicit procedure for designing such controllers is desirable.

The Directional Damping control scheme was then developed for the general case for systems with a single lightly damped mode at a frequency well above the system bandwidth (as is the case for the DTM slide system). The procedure separates the design of the control of the motor/amplifier dynamics from that of the lightly damped slide poles, and produces a controller similar to ZPIC, but with improved disturbance rejection characteristics.

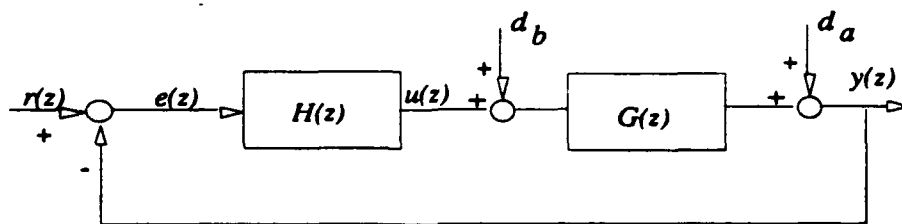
The Zero Placement Integral Control scheme and the effect of the scheme on the surface roughness of diamond turned parts are presented. Also, the derivation of the Directional Damping control scheme is presented followed by the implementation of the scheme on the z-slide of the DTM.

15.2 DTM SLIDE SYSTEM

A schematic block diagram of a slide of the DTM is shown in Figure 1(a). The hydrostatic slide and its coupling to the ball-screw produce a highly underdamped second order system making the system susceptible to disturbance induced vibration. The remaining portion of the system plant $G(s)$ is comprised of the motor and its amplifier which drive the ball-screw. Due to the increased effect of friction in the motor and ball-screw at low feedrates as well as the dynamics of the velocity and current loops within the amplifier, this portion of the system plant exhibits nonlinear dynamics. Although the dynamics of the system change as a function of feedrate, a linear model can be used to describe the dynamics which are representative of the plant in the range of feedrates which are considered to be suitable for precision machining (0 to 10 mm/min).



(a) Schematic Block Diagram



(b) Block Diagram of Discrete System

Figure 1: Block Diagram for a Slide of a Diamond Turning Machine

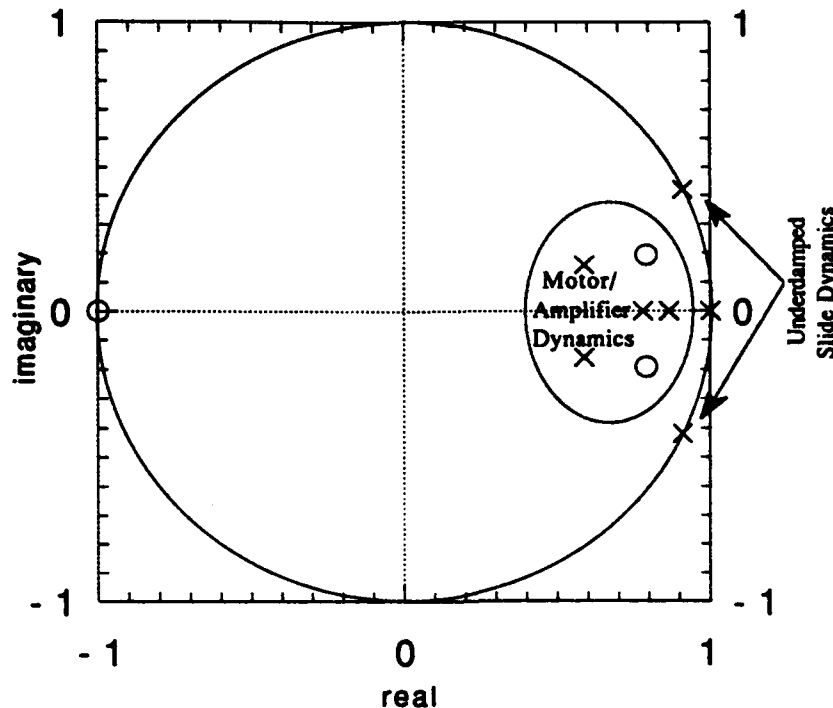


Figure 2: Pole and Zero Locations of the Discrete Representation of the DTM

A discrete version of the seventh order model reported previously [3] will be used to describe the dynamics of the z-slide in this range. A block diagram of this discrete system is shown in Figure 1(b). The pole and zero locations of the DTM ($G(z)$) are shown in Figure 2. The lightly damped poles of the slide which are responsible for the vibration are distinguished from the poles of the motor and amplifier which help determine the system's bandwidth. The goal is to design a controller $H(z)$ which dampens the oscillatory poles while also improving the system bandwidth.

15.3 ZERO PLACEMENT INTEGRAL CONTROL

The design process for the scheme referred to as Zero Placement Integral Control (ZPIC) involves finding (by trial and error) a combination of lead/lag network and complex zero locations which balance the damping of the underdamped poles with the improvement of the system bandwidth (while also incorporating integral control to eliminate steady state following error). Direct cancellation of the lightly damped poles was not effective do to plant uncertainty and the close proximity to the stability boundary. Therefore, the goal of ZPIC is to place complex zeros near the lightly damped poles; but in a more heavily damped region so that for large gains, the poles will approach the zeros making the poles more heavily damped. This is shown in the system's root locus in Figure 3. The bold \times 's and O 's represent the controller poles and zeros. As the gain is

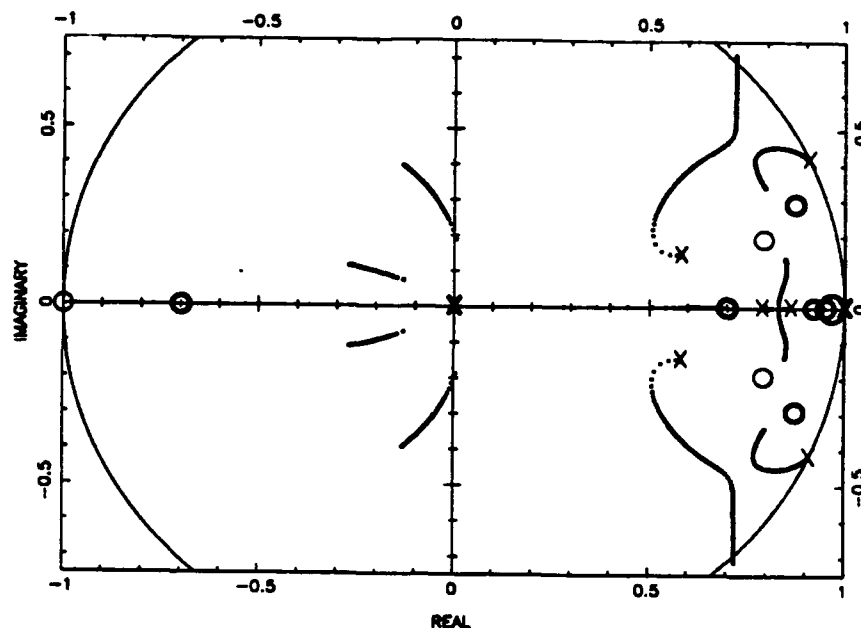


Figure 3: Z Slide Discrete Root Locus using ZPIC

increased, the system poles corresponding to the underdamped slide dynamics move towards the complex pair of controller zeros resulting in a more heavily damped closed loop pole location.

15.3.1 ZPIC Versus PID

The effectiveness of various controllers in rejecting disturbances can be analyzed by viewing the Bode plot of the transfer function from the disturbance input to the position output. Plots of the sensitivity transfer functions for disturbances before and after the plant are shown in Figure 4. Disturbances before the plant include mainly electrical noise from various sources leaking into the control signal. This is a common problem in servo systems with PWM amplifiers. Figure 4(a) shows that PID control will amplify these disturbances in the 60-70 Hz range as its peak reaches approximately 10 db. ZPIC on the other hand peaks below 0 db and will attenuate such disturbances.¹ Disturbances which occur after the plant include such things as acoustic noise and structural machine vibration. Figure 4(b) shows that PID will amplify these disturbances if they are below 65 Hz, and ZPIC will amplify them if they are 80 Hz to 150 Hz. Further improvement of these sensitivity plots will be accomplished in Section 15.4 as the minimization of these plots will be used as design criteria for directional damping control.

¹Much of this electrical noise can be reduced by grounding directly to the earth rather than the building's electrical ground.

The rms amplitude of vibration measured on the z slide is plotted as a function of feedrate in Figure 5(a). ZPIC provides vibration reduction through most of the range of practical machining feedrates with the exception of the feedrates at or near 0 mm/min where the nonlinearity of the friction has the greatest effect. The Fourier transform of the following error to the slide moving at 5 mm/min is shown in Figure 5(b). As suggested by the disturbance sensitivity Bode plots, the large peak under PID control at 62 Hz is reduced significantly by ZPIC.

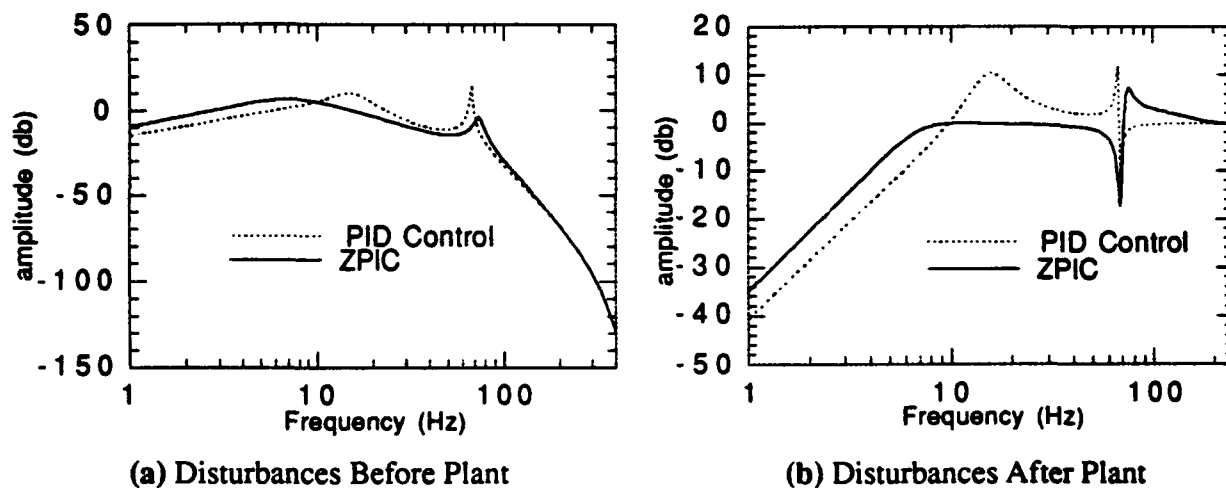


Figure 4: Frequency Responses of Sensitivity Transfer Functions

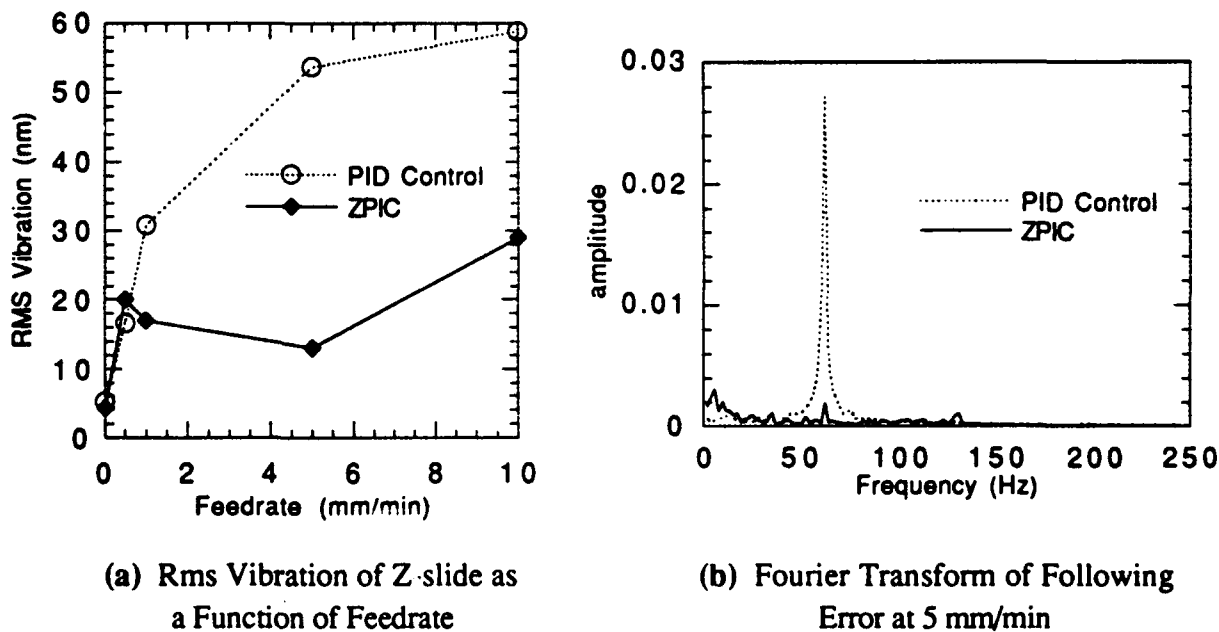


Figure 5: Experimental Z-slide Vibration

15.3.2 Effect on Surface Finish

ZPIC has been shown effective in reducing the level of slide vibration. However, a more important concern for diamond turning, is how that vibration translates into surface finish. Samples of OFHC copper were machined while controlling the z slide with ZPIC and PID control to determine the effect of each control on surface finish.

As can be seen in Figure 5(a), the level of vibration is significantly higher when the slides are moving. Thus, the effect of the vibration will be most severe on parts involving the motion of both axis. Therefore, sections of cones were machined to best demonstrate the effect of control on surface finish. The parts were machined with the x and z slides moving at 8 mm/min and 5 mm/min respectively and a spindle speed of 1000 RPM.

The surfaces machined with each type of control are shown in Figures 6 and 7, with the measured vibration of the z slide and the resulting surface finish summarized in Table 1. The improved level of vibration of ZPIC produced a significantly better surface finish on the machined parts, reducing the surface roughness from 264 nm PV to 98 nm PV (measured over an $780\text{ }\mu\text{m} \times 570\text{ }\mu\text{m}$ area).

	PID		ZPIC	
	RMS (nm)	PV (nm)	RMS (nm)	PV (nm)
Measured Z-slide Vibration	54	200	13	90
Surface Finish	49	264	17	98

Table 1: Slide Vibration and Corresponding Surface Finish for OFHC Copper Machined using PID Control and ZPIC

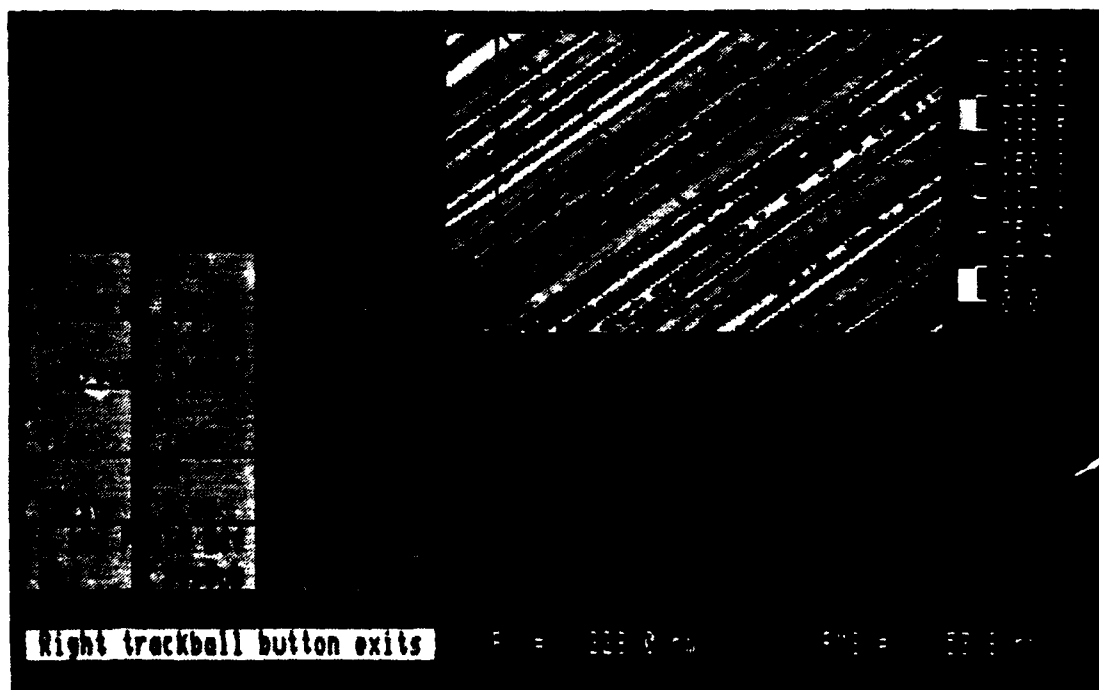
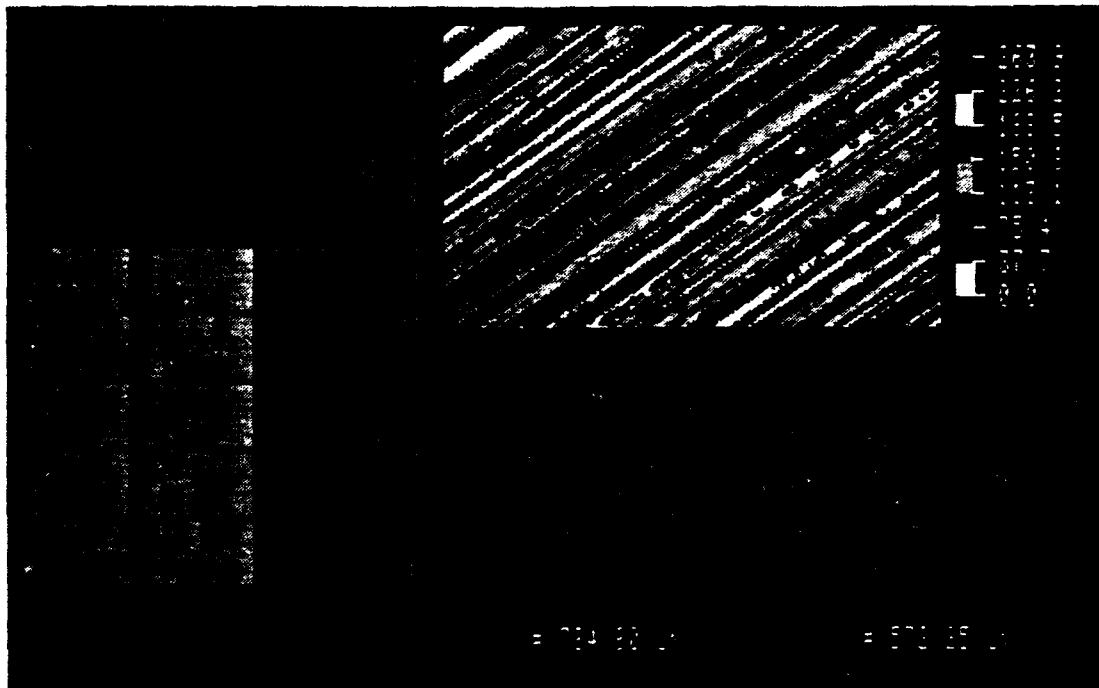


Figure 6: Surface of OFHC Copper Machined with PID Control

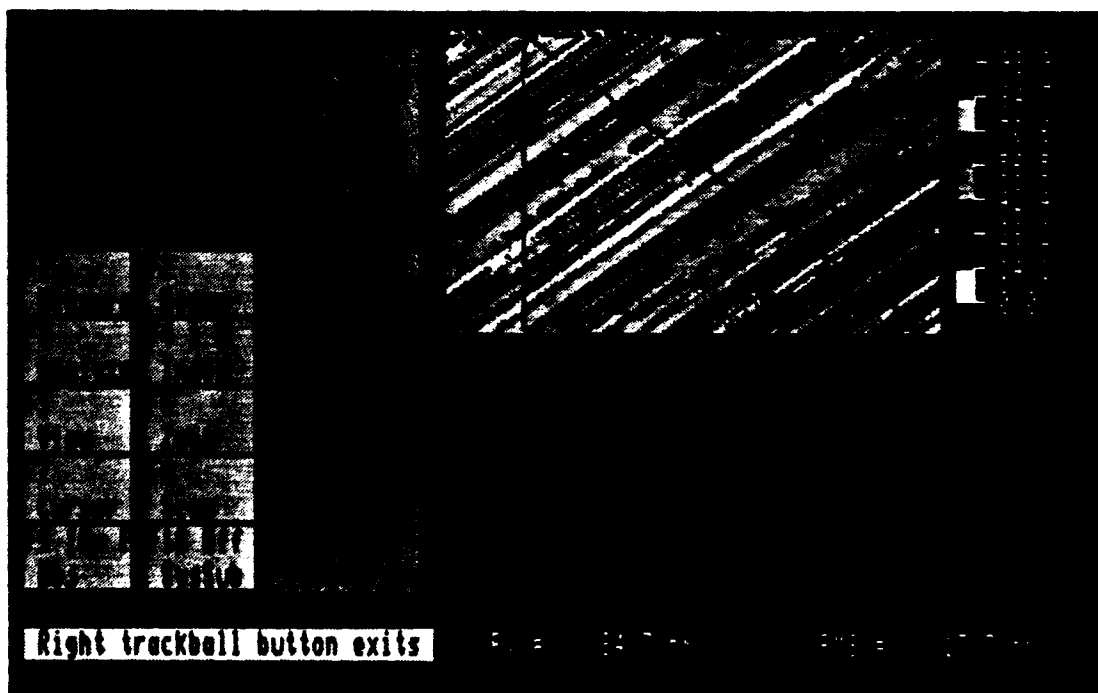
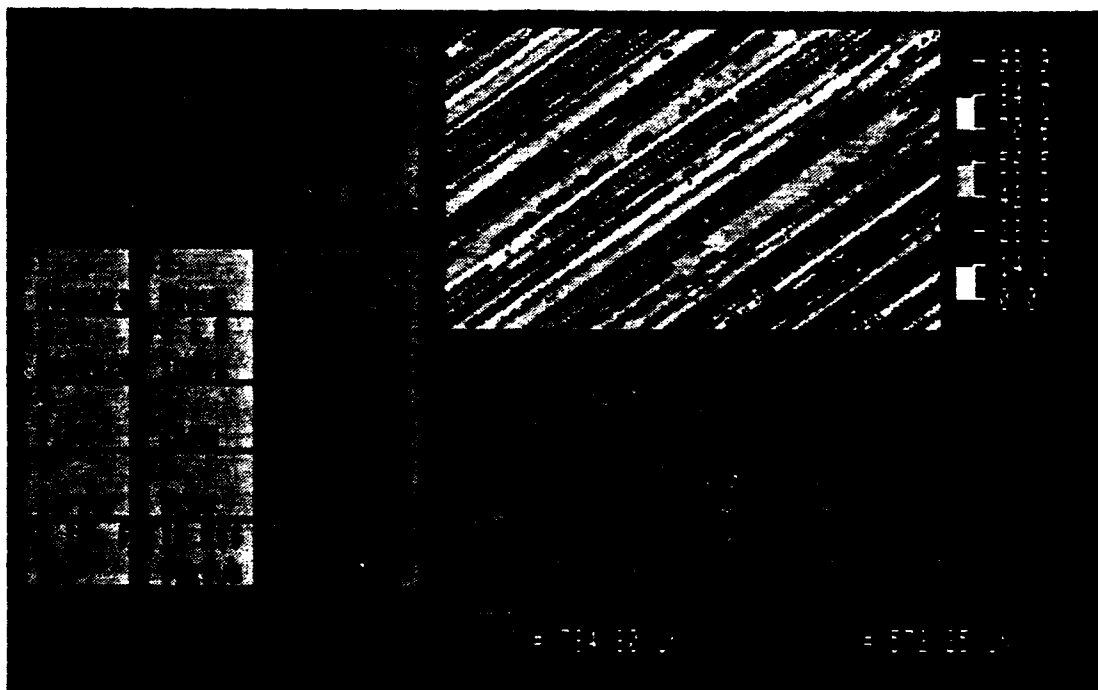


Figure 7: Surface of OFHC Copper Machined with ZPIC Control

15.4 DIRECTIONAL DAMPING CONTROL

ZPIC has proven to be effective in reducing the vibration of the z-slide (thus improving surface finish) while also maintaining an acceptable system bandwidth. However, the design of the scheme was quite heuristic. To be applicable to other similar systems with different dynamics or to optimize the design for this particular system, a more rigorous set of design guidelines and criteria is necessary.

In this section a design procedure is presented for systems with characteristics similar to that of the DTM slide. Such systems have a highly underdamped mode coupled to a set of dynamics which determine the system's bandwidth. The portion of the control which determines the system bandwidth is designed independently of the lightly damped mode. Subsequently, the incorporation of a complex pair of zeros minimizes the sensitivity to disturbances without disrupting the system bandwidth. For the DTM slide system, this procedure separates the design of the motor/amplifier dynamics from that of the lightly damped slide poles. The result is a scheme - Directional Damping Control (DDC) - which is similar to ZPIC, but with improved disturbance rejection characteristics.

15.4.1 Design Procedure

Problem Definition It is assumed that the discrete time transfer function of the N^{th} order system to be controlled can be expressed as

$$G(z) = \frac{K_G \prod_{i=1}^N (z - a_i)}{(z^2 + 2\sigma z + \sigma^2 + \omega^2) \prod_{j=1}^{N-2} (z - b_j)} \quad (1)$$

where $z^2 - 2\sigma z + \sigma^2 + \omega^2$ is the lightly damped pair of complex poles at $z = \sigma \pm i\omega$, which make the system susceptible to vibration. The remaining poles and zeros are more heavily damped, and one or more of these poles play a major role in determining the system's bandwidth. The objective is to dampen the oscillatory poles while achieving the desired system bandwidth and time response.

Design Strategy In the development of ZPIC, a pair of complex zeros was placed near the lightly damped poles, but in a more heavily damped region. This was done keeping in mind that as the controller gain goes to infinity, the poles will move to the zeros; thus at high gains the poles will tend to be in a more damped region. However, it was found that the gains which are implementable (input saturation restricts the gain) are much less than those which make the poles

approach the zeros. Therefore, in this scheme, rather than placing the zeros to attract the poles at high gains, the zeros will be used to direct the poles to the damped region at low gains.

Because the zeros are not used to cancel the poles, the scheme will not be sensitive to variations or uncertainty in the location of the lightly damped poles (assuming the distance between the poles and the zeros is several times the uncertainty of the pole location). However, the distance between the zeros and the lightly damped poles will be small compared to the distance to the rest of the system poles. This will have the effect of making the rest of the system dynamics insensitive to the exact location of the complex zeros; because for poles far away, the effect of the complex zeros will cancel the effect of the lightly damped poles.

Since these lightly damped poles and zeros will have little effect on the rest of the system, design can first be carried out on the rest of the system ignoring the oscillatory poles. Once the controller has been designed to produce the desired system bandwidth and steady state characteristics, then the complex pair of zeros can be added to achieve the vibration reduction. This procedure is formalized below.

Formalization The following steps are used in designing a discrete time directional damping controller (DDC) utilizing root locus and frequency response techniques [4,5].

- I. *Modify system $G(z)$ by replacing the lightly damped poles with two poles at the origin to form $G_{mod}(z)$.*

$$G_{mod}(z) = \frac{K_G \prod_{i=1}^N (z - a_i)}{z^2 \prod_{j=1}^{N-2} (z - b_j)} \quad (2)$$

These poles are at the origin in $G_{mod}(z)$ because a pair of poles will be placed at the origin when the complex zeros are added.

- II. *Design forward loop controller $H_{mod}(z)$ for the modified system $G_{mod}(z)$ to achieve desired system bandwidth and steady state characteristics.* In this step, a lead-lag network can be used to increase the system bandwidth and assure that system and controller poles remain relatively far from the location of the lightly damped poles. Also, if desired, integral control can be incorporated to eliminate steady state following error.

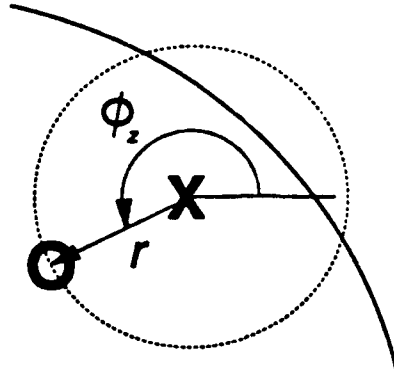


Figure 8: Complex Zero Placement Relative to Lightly Damped Pole

- III. *Add complex pair of zeros at the angle ϕ_z (defined in Figure 8) that will give the lightly damped poles the desired angle of departure to minimize sensitivity to disturbances. The root locus angle criterion is used to determine the ϕ_z needed to produce the specified angle of departure. This criterion requires that for a system with transfer function $T(z)$ the root locus can exist only at the points z which satisfy*

$$T(z) = n \cdot 180^\circ, \quad n = \pm 1, \pm 3, \pm 5, \dots$$

Applying this criterion near one of the lightly damped poles at a desired angle of departure ϕ_d and using $n = 1$, the needed angular position of the complex zero is

$$\phi_z = 180^\circ - \phi_d - \angle G_{mod}(\sigma + i\omega) H_{mod}(\sigma + i\omega) \quad (3)$$

In obtaining this result, the respective conjugates of the lightly damped pole and zero are ignored since they are at approximately the same angle with the pole, causing their effects to cancel one another. The complex zeros are then placed at

$$\alpha_{1,2} = (\sigma + r \cos \phi_z) \pm i(\omega + r \sin \phi_z) \quad (4)$$

forming the controller

$$H(z) = \frac{(z - \alpha_1)(z - \alpha_2)}{z^2} H_{mod}(z) \quad (5)$$

To minimize the sensitivity to disturbances, ϕ_d should be chosen to place the closed-loop pole location in direct line with the origin. This can be done by first choosing the angle of departure to be directly towards the center of the unit circle,

$$\phi_z = (2f_n T + 1)180^\circ \quad (6)$$

f_n = natural frequency of slide (Hz)

T = sampling rate (s)

and then modifying ϕ_d by the angle which the pole location deviates from the line to the origin.

The distance r between the poles and zeros should be several times larger than the uncertainty of the pole location to assure insensitivity to parameter variation and modeling error. However, if r is chosen too large the system bandwidth could be effected. The values of r which balance the effects on disturbance rejection and bandwidth can be determined by viewing the disturbance sensitivity frequency response and the closed loop frequency response plots for various values of r .

15.4.2 Implementation on the DTM

The preceding procedure was carried out for the z-slide of the DTM. This process will not be discussed in detail here, but can be found in [2]. A PID controller combined with a lead-lag network is used for H_{mod} to improve the bandwidth of the system and eliminate the steady state following error. An angle of departure of $\phi_d = 195^\circ$ with a pole-zero distance $r = .07$ is specified to give the proper complex controller zero location. The system root locus using this control is shown in Figure 9. Note that to achieve the desired angle of departure, the directional damping zeros are required to be non-minimum phase zeros. This does not cause any stability problems since it would take much higher gains than will be used for the poles to track to the zeros. Rather, it is this positioning of the zeros which results in the motion of the poles to the most damped region.

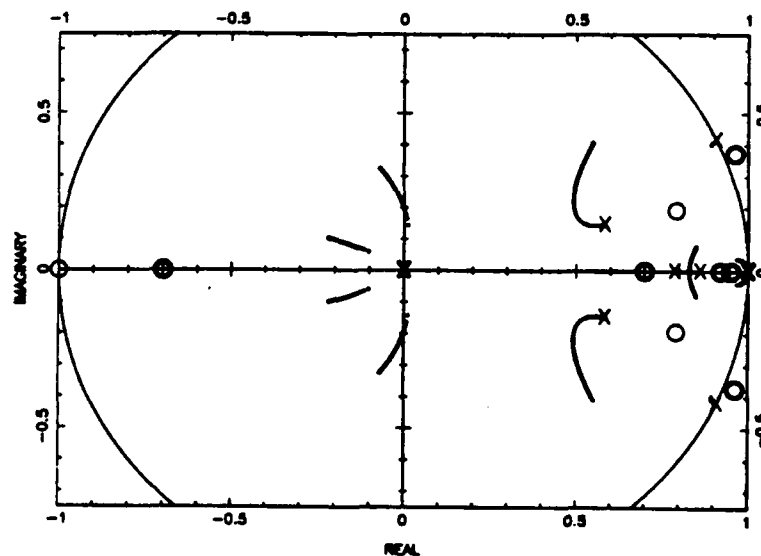
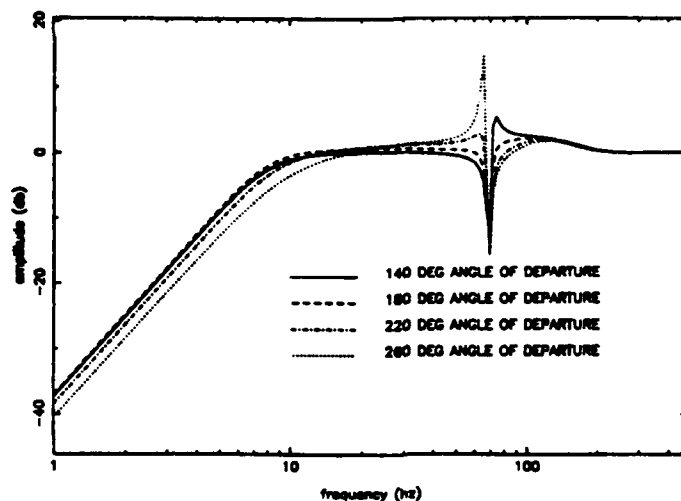
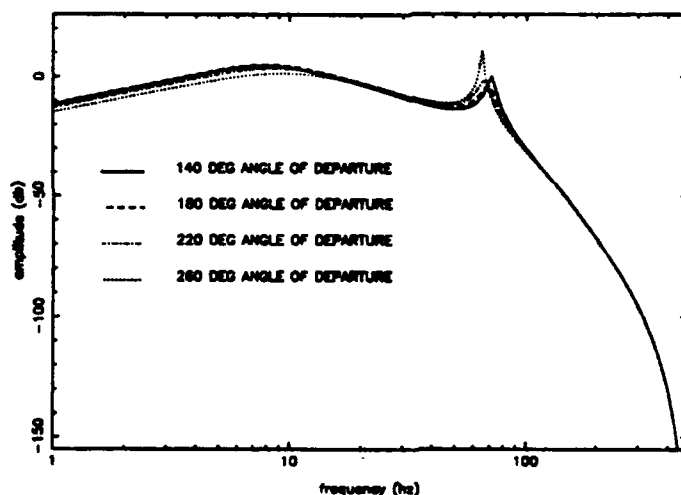


Figure 9: Z-slide Root Locus using Directional Damping Control



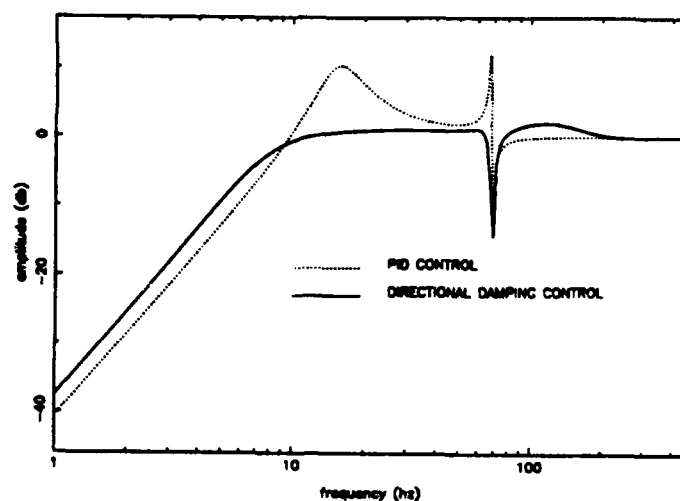
(a) Disturbance After the Plant



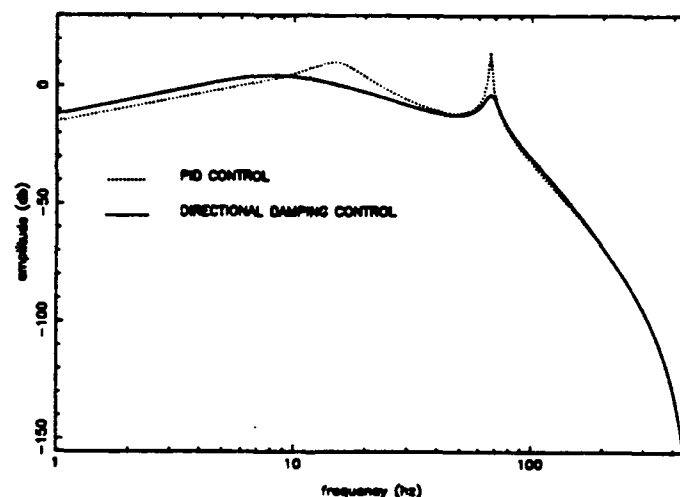
(b) Disturbance Before the Plant

Figure 10: Sensitivity Frequency Responses for DDC Using Various Angles of Departure

To illustrate the benefit of specifying the angle of departure using DDC, the effect of that angle of departure (and ultimately the closed loop pole location) on the sensitivity to disturbances is shown in Figure 10. For angles of departure that are less than the design value ($\phi_d = 195^\circ$), the closed loop poles will be at a higher frequency and less damped. If the angle of departure is chosen to be greater than the design value, less damped lower frequency poles will result. This causes a peak in sensitivity at frequency just above the open loop natural frequency for reasonably smaller angles of departure, and just below the open loop natural frequency for larger angles of departure.



(a) Disturbances After the Plant



(b) Disturbances Before the Plant

Figure 11: Comparison of Sensitivity Frequency Responses for DDC and PID Control

The disturbance sensitivity frequency responses for DDC using the design values $\phi_d = 195^\circ$ and $r = .07$ are compared to that of PID control in Figure 11 showing a significant reduction in the sensitivity to disturbances with DDC. The following error to a 5 mm/min ramp input is compared in Figure 12. The design of the directional damping controller has produced a controller that is quite similar to ZPIC discussed in Section 15.3; with the exception that by controlling the angle of departure of the highly underdamped poles, greater reduction of the disturbance sensitivity transfer functions was achieved with DDC.

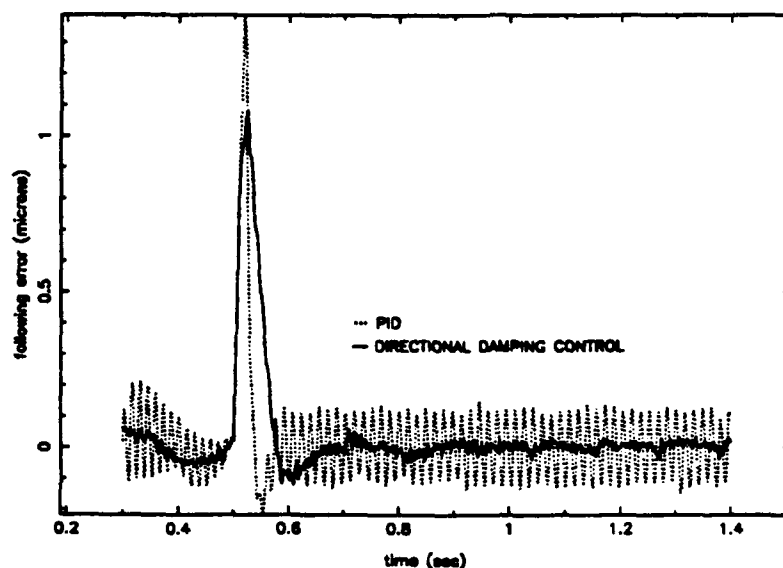


Figure 12: Comparison of Following Error to a 5 mm/min Feed for DDC and PID Control

15.5 CONCLUSION

Previous analysis of various existing control schemes identified the major issue in the design of a control scheme for the DTM to be the need to dampen the highly underdamped dynamics of the hydrostatic slide, while also improving (or maintaining) the system bandwidth. Directional Damping Control was developed as a way to separate the design of the bandwidth determining dynamics from the underdamped slide dynamics. In doing so, the system's sensitivity to disturbances was greatly reduced without sacrificing the system bandwidth. The reduction in the level of slide vibration produced a corresponding improvement in the surface finish of machined parts.

While the scheme designed improved the level of vibration throughout most of the range of practical machining feedrates, the reduction of vibration was not uniform throughout the range (particularly at the lowest feedrates). The "representative" linear model of the system used in the design, is much less "representative" of the nonlinear system as the feedrates approach zero. The incorporation of directional damping control into an adaptive scheme may provide improved performance at all feedrates.

References

- [1] Abler, J. A., Ro, P. I., "Slide Motion Control Schemes for Vibration Reduction," *Precision Engineering Center 1991 Interim Report*, pp. 29-36.
- [2] Abler, J. A., "Control of Precision Slide Motion for Vibration Reduction in Diamond Turning," *MS Thesis*, North Carolina State University, 1991.
- [3] Abler, J. A., Ro, P. I., "Slide Motion Control Algorithms for Vibration Reduction," *Precision Engineering Center 1990 Annual Report*, Vol. VIII, pp. 285-306.
- [4] Evans, W. R., "Control System Synthesis by Root Locus Method," *AIEE Transactions*, vol. 69, part 1, 1950, pp. 66-69.
- [5] Palm, W. J., *Modeling, Analysis and Control of Dynamic Systems*, Wiley & Sons, 1983.

16 ENHANCEMENTS TO THREE AXIS DTM CONTROLLER

Michele H. Miller

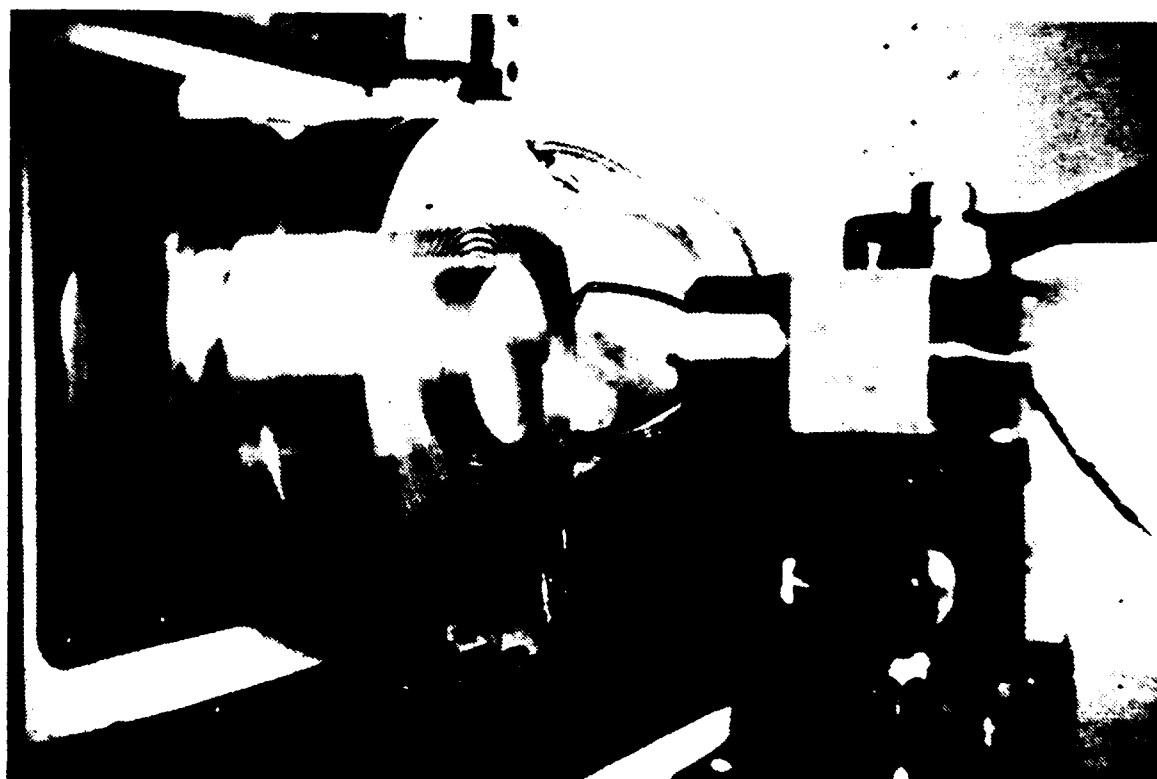
Graduate Student

Thomas A. Dow

Professor

Department of Mechanical and Aerospace Engineering

Several functional improvements have been made to the three axis diamond turning machine controller. Compensation schemes for slideway errors have been implemented and tested. The fast tool servo controller was integrated with the axes controller. With this new system an off-axis parabolic section was machined with 1.1 wave PV accuracy over a 114 mm aperture. The tasks required for machining this surface are described in this report.



16.1 INTRODUCTION

The capability of current diamond turning machines is limited by the ability of the machine tool controller. Commercially available controllers are neither fast enough nor flexible enough to incorporate new concepts such as fast tool servos or real-time geometric error correction. As a result, the capabilities for non-rotationally symmetric surfaces and improved surface finish and figure accuracy have not been fully realized.

The reasons for this failure are the large volume of data and the high speed at which it must be utilized. For the case of spherical and aspherical surfaces which are rotationally symmetric, the amount of data required is a function of the radius of the part, the feedrate, and the speed of the control algorithm. The data requirements increase enormously for non-rotationally symmetric parts achievable with a fast tool servo because the entire surface area must be defined rather than just a radial cross-section. For real-time correction of repeatable slideway errors, each axis must know the location of the other axis and a table lookup or error computation scheme implemented. Most computer architectures are not designed to accommodate both the high-speed processing and large data transfer rates of these applications.

A multiprocessor architecture has been designed to overcome these shortcomings (see Sections 17 and 18). It provides both the computational speed and communications and I/O interfaces necessary to completely control the machine for rotationally or non-rotationally symmetric surfaces. Integration of the fast tool servo and correction for repeatable geometric errors have been demonstrated, and the software algorithms which accomplish these applications are described.

16.2 REPEATABLE SLIDEWAY ERROR COMPENSATION

16.2.1 X and Z Straightness Errors

One of the enhancements to the DTM controller has been the addition of software to correct repeatable slideway errors. The X and Z straightness errors were mapped. Based on error tables, X straightness errors are compensated by the Z slide, and Z straightness errors are compensated by the X slide. The effectiveness of this scheme was determined by performing straightness measurements with and without the correction. Figures 1(a) and (b) show the results: the X straightness improved from 200 nm to 40 nm PV error while the Z slide straightness improved from 60 nm to 40 nm PV. The repeatability of the straightness measurements is approximately 40 nm.

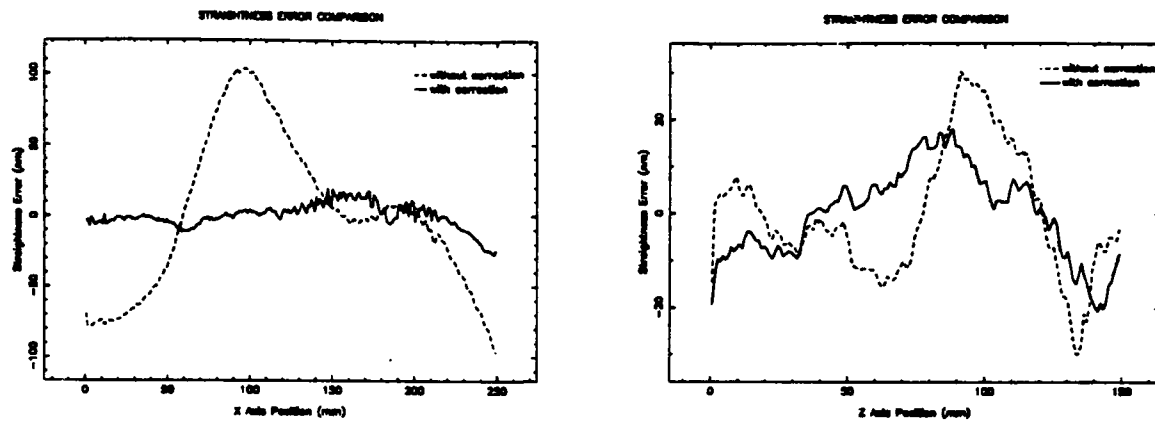


Figure 1: (a) X (b) Z straightness errors with and without controller compensation (note the difference in scale between the two plots)

16.2.2 Z Slide Yaw Error

The angle of the spindle axis with respect to the X axis is another source of error. Any deviation from perpendicular causes a cone to be superimposed onto the machined surface. In addition, Z slide yawing causes an error in the slide's position measurement because of an Abbe offset (the interferometer retroreflector is offset from the center of the Z axis by 166 mm). These two errors are depicted in Figure 2.

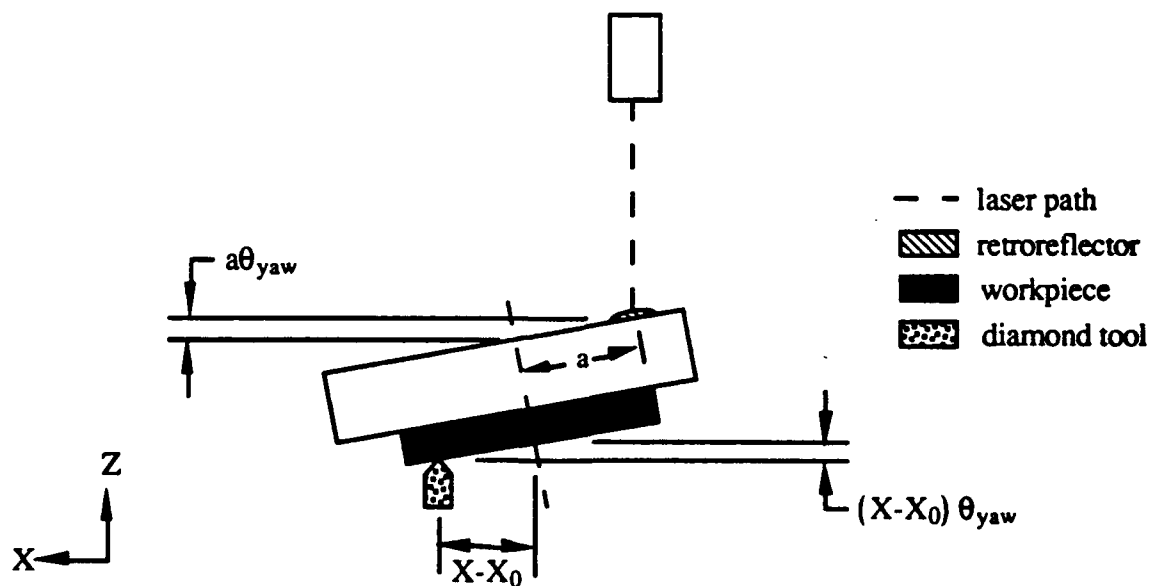


Figure 2: Z direction errors caused by Z axis yaw

Measurements have shown that θ_{yaw} varies nearly linearly with Z position and can be modelled by an equation of the form:

$$\theta_{yaw} = m Z + b \quad (1)$$

Because the Z slide yaws 1.6 arcsec over its range of 150 mm, the slope of Equation (1) is

$$m = \frac{1.6 \text{ arcsec}}{150 \text{ mm}}$$

The intercept b is equal to θ_{yaw} at $Z=0$ (where the slide is fully extended). To compensate for θ_{yaw} , the correction algorithm modifies the Z reference point by,

$$(\Delta Z)_{ref} = (X - X_0) \theta_{yaw} + a \theta_{yaw} \quad (2)$$

where θ_{yaw} is represented in radians. The *change* in θ_{yaw} is repeatable. Its absolute value at any Z position, however, is not repeatable because the screws attaching the spindle and motor assembly to the Z slide allow for adjustment. The correction algorithm takes this variability into consideration by allowing the user to input the yaw angle at a given Z slide position. Given a θ and Z, the algorithm calculates b, the intercept of Equation (1). Then in real-time θ_{yaw} and $(\Delta Z)_{ref}$ are calculated based on current X and Z positions during machining.

To test the algorithm a flat was machined without yaw correction while the Z slide was positioned at $Z=86$ mm. Based on the height of cone machined onto the flat, the yaw angle was estimated to be -0.58 arcsec. The values of -0.58 and 86 mm were entered to reset the intercept of Equation (1):

$$b = \theta_{yaw} - m Z \quad (3)$$

$$b = -0.58 \text{ arcsec} - \frac{1.6 \text{ arcsec}}{150 \text{ mm}} (86 \text{ mm})$$

$$b = -1.50 \text{ arcsec}$$

Flats subsequently machined at $Z=86$ mm and $Z=16$ mm showed that the correction did indeed remove the slight cone shape. Figure 3 shows how the amount of corrective action taken by the Z slide changes depending on its absolute position. For three nominal Z positions ($Z=16$, 66, and 116 mm), data was collected while the X slide traversed from $X-X_0 = -39$ mm to $X-X_0 = 0$ and the Z slide "held position." The slope of each line is the yaw angle at that Z position. The correction is largest for the $Z=16$ mm case where the yaw angle calculated by the algorithm is,

$$\theta_{yaw} = \frac{1.6 \text{ arcsec}}{150 \text{ mm}} Z - 1.50 \text{ arcsec}$$

$$\theta_{yaw} = -1.33 \text{ arcsec}$$

The amount of correction at $X-X_0 = 0$ is due to the Abbe offset. When machining flat surfaces, this offset is constant and introduces no form error. However, for surfaces requiring Z slide motion, the Abbe offset would introduce figure error if left uncorrected.

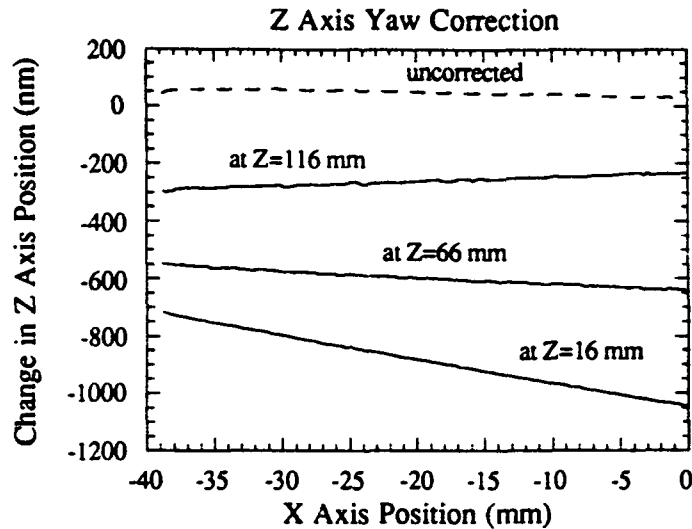


Figure 3: DTM slide data showing yaw correction

16.3 FAST TOOL SERVO MACHINING IMPROVEMENTS

In the past year control of the fast tool servo has been fully integrated with the X and Z axis control. As shown in Figure 4, the fast tool servo (the Z' axis) attaches to the X slide and provides motion in the Z direction. Non-rotationally symmetric surfaces are described by $Z=Z(\rho, \theta)$. The X slide position corresponds to the ρ coordinate; the angular position of the spindle corresponds to the θ coordinate; and the Z' and Z axes combine to produce the Z dimension.

For the machining experiments described in the *1990 Annual Report* the FTS controller operated *independently* from the 68020 based X and Z axis controller. That FTS controller calculated reference points based on angular position of the workpiece (provided by spindle encoder) and an estimated X axis position (calculated by counting the number of spindle revolutions, multiplying by the feedrate, and dividing by the spindle speed). The new *integrated* FTS controller has access to the laser interferometers and calculates reference points based on the actual X position rather than an estimated position. With the new architecture an off-axis parabolic section (f number =

1079.5 mm, $X_0 = 296.47$ mm, workpiece diameter = 127 mm) was machined on-axis. The 1990 report described the pre-processing which was necessary to minimize the amount of real-time computation for generating several types of non-axisymmetric surfaces. Machining off-axis conics on-axis also requires some pre-processing to transform from the coordinate system of the parent conic to the machine coordinates used to generate the off-axis section. Finding the optimum tilt angle between these two coordinate systems is discussed below. Machining a measurable surface such as a parabolic section has placed new emphasis on eliminating FTS error sources. Two of these sources--tool radius error and dynamic positioning error--are also discussed.

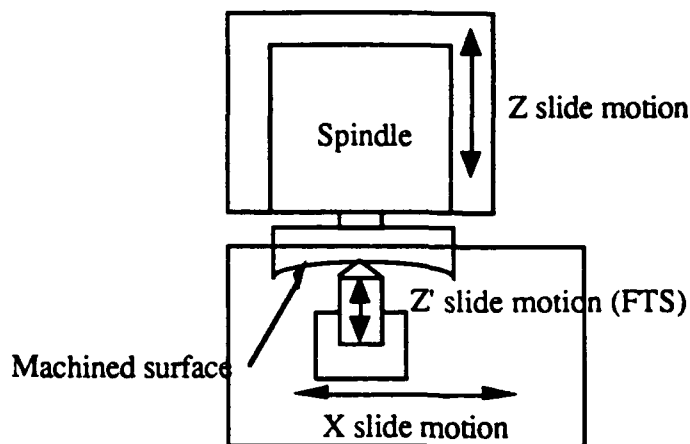


Figure 4: Schematic of machine axes

16.3.1 Tilt Angle Optimization for Machining an Off-Axis Conic Section

Machining an off-axis conic section on-axis requires a transformation from the coordinate system of the parent conic to a local coordinate system of the off-axis segment. These two coordinate systems are shown in Figure 5. The angle of rotation between the two systems should be chosen to minimize fast tool servo motion. By doing so, a wider range of off-axis surfaces may be machined.

According to Gerchman [1], non-axisymmetry is minimized when the surface is tilted so that the points $(\rho, \theta) = (\rho_{\max}, 0^\circ)$ and $(\rho_{\max}, 180^\circ)$ are at the same height. Figure 6(a) shows a cross-section of an off-axis segment through the $\theta=0^\circ/180^\circ$ plane. The equation of the parent surface, the center point of the off-axis segment (X_0, Z_0) , and the workpiece diameter d_w define the section to be machined. The problem, then, is to find the tilt angle α which levels points A($\theta=0^\circ$) and B($\theta=180^\circ$). An analytical solution is difficult since it involves solving a set of nonlinear algebraic equations. There is, however, a simple numerical approach. As a first approximation α can be chosen to correspond to the slope of the conic at X_0, Z_0 :

$$\alpha_0 = \left[\tan^{-1} \left(\frac{dz}{dx} \right) \right]_{X_0, Z_0} \quad (4)$$

Figure 6(b) shows the off-axis section in the coordinate system rotated by α_0 .

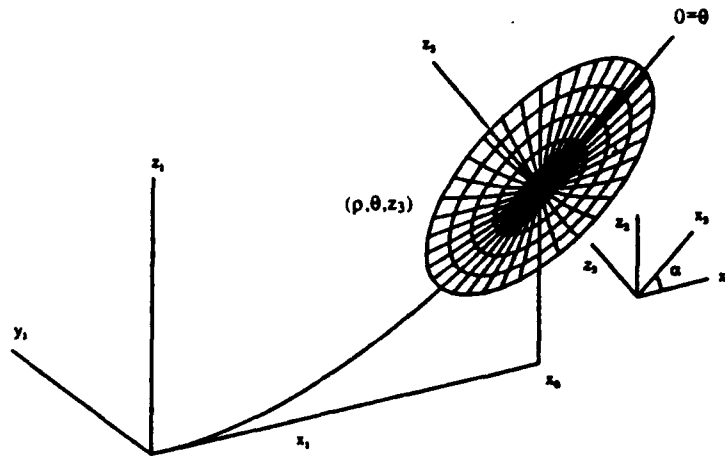


Figure 5: Relevant coordinate systems for machining an off-axis conic section:
(1) the X, Z system of the parent conic; (2) the ρ, Z, θ machine coordinate system

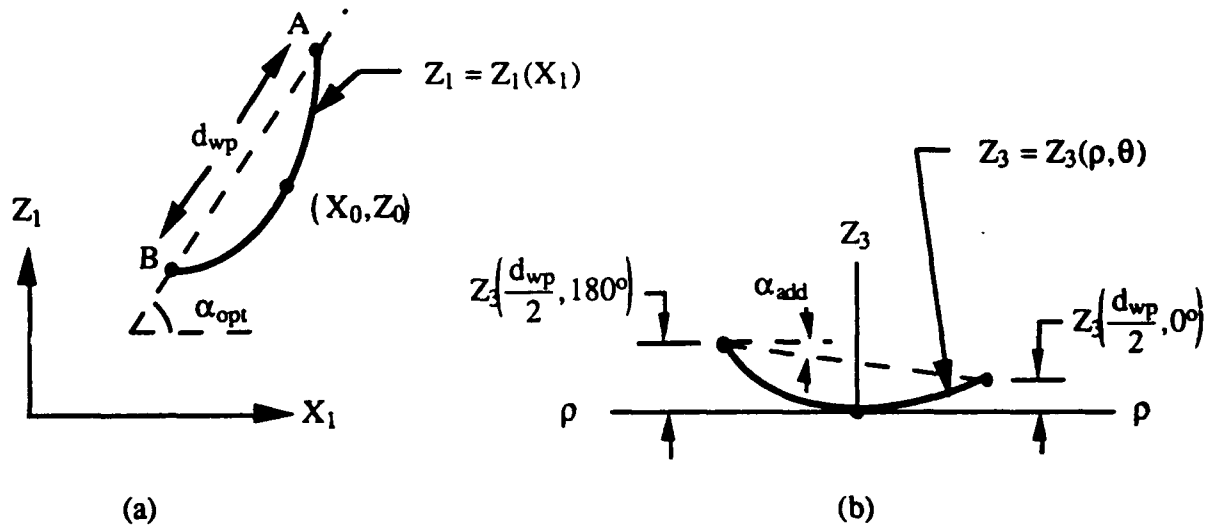


Figure 6: (a) Cross-section of off-axis section in the coordinate system of the parent conic; (b) cross-section of same section in transformed coordinate system (rotated by α_0)

The tilt optimization procedure was carried out for the parabolic section described above. Figure 7(a) shows the Z motion for one revolution at the OD of the segment ($p = 63.5$ mm where the non-axisymmetry is a maximum) for the workpiece tilted by α_0 . Note the difference in Z values for θ

$= 0^\circ$ and $\theta = 180^\circ$. This difference divided by the part diameter is the amount of additional tilt required:

$$\tan \alpha_{\text{add}} = \frac{Z\left(\frac{d_{\text{wp}}}{2}, 0^\circ\right) - Z\left(\frac{d_{\text{wp}}}{2}, 180^\circ\right)}{d_{\text{wp}}} \quad (5)$$

The tilt angle then becomes:

$$\alpha_1 = \alpha_0 + \alpha_{\text{add}} \quad (6)$$

With a new coordinate system rotated by α_1 , the points $(d_{\text{wp}}/2, 0^\circ)$ and $(d_{\text{wp}}/2, 180^\circ)$ shift slightly from their locations in the system rotated by α_0 . Therefore, α_1 is not exactly optimum. By iterating on Equation (5), the optimum angle can be approached. For the example given, two iterations yields a tilt angle $\alpha_2 = \alpha_1 + \alpha_{\text{add}}$ which levels $(d/2, 0^\circ)$ and $(d/2, 180^\circ)$ to within 2 nm. Figure 7(b) shows the non-axisymmetric motion at $p=63.5$ mm when the workpiece has been tilted by α_2 .

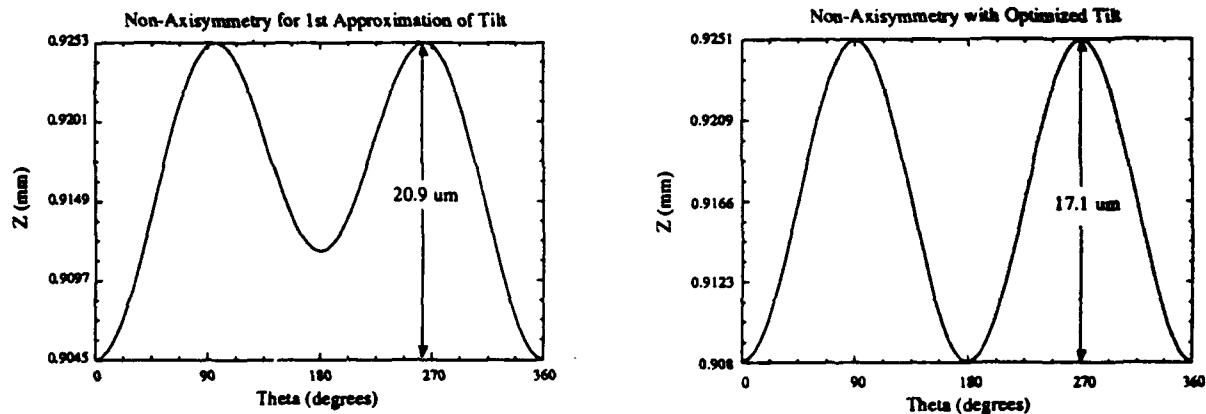


Figure 7: (a) Z motion with non-optimized tilt angle $\alpha_0 = 0.1364642$ radians; (b) Z motion with optimized tilt angle $\alpha_2 = 0.13640805$ radians

As shown in Figure 7(a) and (b) the range of FTS motion reduces from 20.9 μm to 17.1 μm when the tilt is optimized.

16.3.2 FTS Motion Error Due to Tool Radius

When turning contours without a B axis the diamond tool cannot be considered a single point because the actual cutting point moves along the edge of the tool. The tool radius compensation for

turning axisymmetric surfaces has been described in [2]. Implementing an FTS introduces a new tool radius error whose source is described in Figure 8.

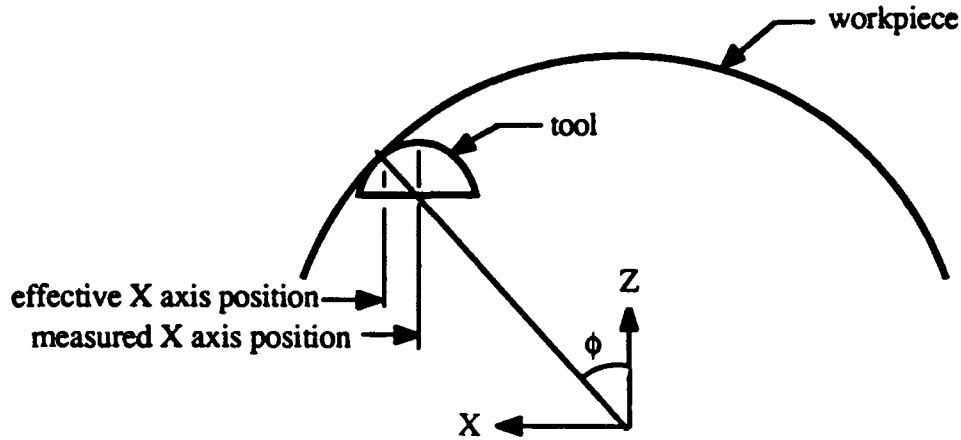


Figure 8: Tool/workpiece interface

The FTS reference point is calculated based on the X axis position provided by the laser interferometer and referenced to the centerline of the spindle. As Figure 8 shows, the cutting point on the tool will be outside that position. The FTS reference point should be calculated based on the X coordinate of the cutting point. The error in the Z direction is then

$$e_Z = f(X_{\text{cut}}, \theta) - f(X_{\text{laser}}, \theta) \quad (7)$$

where $f(X, \theta)$ is a function describing only FTS motion.

The size of this error depends on the shape of the workpiece and dimensions of the tool. The error for the test workpiece in combination with two different tools has been calculated. Tool #1 (radius $r=0.762$ mm, clearance angle $\alpha=6^\circ$) was the tool used to machine the off axis parabola described above in September 1991. Tool #2 ($r=2.54$ mm, $\alpha=10^\circ$) is a larger radius tool which should produce a better surface finish. The larger radius and clearance angle of this tool will cause a bigger error. The best fit asphere to this off-axis parabola is very close to a sphere of 2180 mm radius. For a sphere, the difference between X_{cut} and X_{laser} is:

$$\delta X = |X_{\text{cut}} - X_{\text{laser}}| = \frac{b}{\sqrt{1 + \frac{a^2}{b^2} \tan^2(90^\circ - \phi)}} \quad (8)$$

where,

a = major axis dimension of tool ellipse = $r/\cos\alpha$

b = minor axis dimension of tool ellipse = r

ϕ = angular distance from center of workpiece

At the outside edge of the workpiece δX will be a maximum, and because the slope of the FTS motion, $\delta Z/\delta X$, is also a maximum, the figure error is calculated at that point. For a 127 mm diameter part with 2180 mm radius of curvature,

$$\tan \phi_{\max} = \frac{127/2}{2180}$$

$$\phi_{\max} = 1.67^\circ$$

For tool #1:

$$a = 0.762/\cos 6^\circ = 0.7661973 \text{ mm}$$

$$b = 0.762 \text{ mm}$$

$$\delta X = 0.022085 \text{ mm}$$

For tool #2:

$$a = 2.54/\cos 10^\circ = 2.5791836 \text{ mm}$$

$$b = 2.54 \text{ mm}$$

$$\delta X = 0.072899 \text{ mm}$$

The slope of the FTS motion with respect to the radial direction is greatest at (X, θ) equals (63.5, 0). Figure 9 shows FTS motion at $\theta = 0, 45^\circ, 90^\circ, 180^\circ$. (At 90° and 270° the FTS motion is zero for all X .)

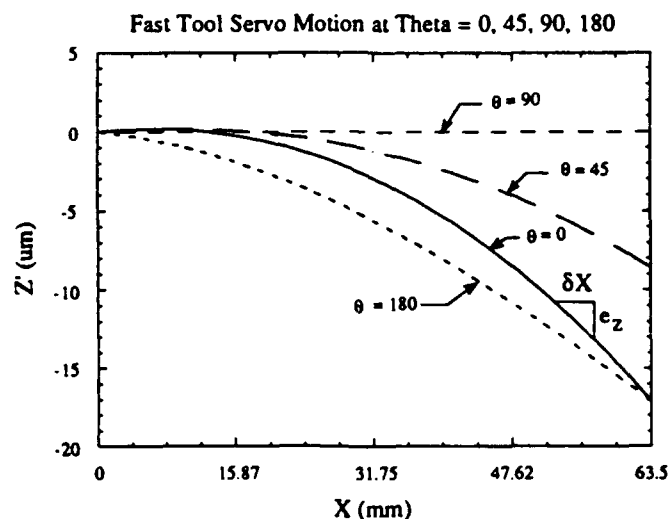


Figure 9: Fast tool servo motion at four angular cross-sections

For tool #1 when $X_{\text{cut}} = 63.5$ mm, $X_{\text{laser}} = X_{\text{cut}} - \delta X = 63.477915$ mm. For this case,

$$\begin{aligned} e_z &= f(X_{\text{laser}}, \theta) - f(X_{\text{cut}}, \theta) \\ &= 16.581 \mu\text{m} - 16.596 \mu\text{m} \\ &= -0.015 \mu\text{m} \end{aligned}$$

For tool #2 when $X_{\text{cut}} = 63.5$ mm, $X_{\text{laser}} = 63.427101$ mm. The bigger δX causes a bigger error:

$$\begin{aligned} e_z &= 16.548 \mu\text{m} - 16.596 \mu\text{m} \\ &= -0.048 \mu\text{m} \end{aligned}$$

This is the amount of error in the Z direction. The figure error perpendicular to the surface is:

$$e_z \cos \phi = e_z \cos(1.67^\circ) = 0.9996 e_z$$

or nearly the same as the Z direction error for this particular surface.

To compensate for this error, Equation (8) must be evaluated during each control cycle and the FTS reference point calculated based on X_{cut} (instead of X_{laser}). Tool radius compensations for axisymmetric surfaces can be made off-line when reference points are generated or in real-time at the cycle time of the X and Z axis controller. Because FTS reference points are not generated off-line and because the FTS control rate is much higher than the X and Z axis control, the tool radius error described above is more difficult to compensate. A fast way to evaluate or approximate Equation (8) is necessary.

16.3.3 Closed Loop Control of FTS

In the *1990 Annual Report* the FTS machining described was performed open loop. With its high bandwidth the dynamic response of the FTS is excellent. However, due to piezoelectric hysteresis, positioning accuracy is poor. Figure 10 shows FTS extension versus applied voltage. To improve positioning accuracy, feedback control of the servo was added to the DSP control program. The algorithm implemented was integral control with effort limiting used previously by Falter [3]. Figure 11(a) shows a block diagram of the control system. The elements of the controller are an integral control effort calculation and a filter which limits the control effort. These are shown in Figure 11(b).

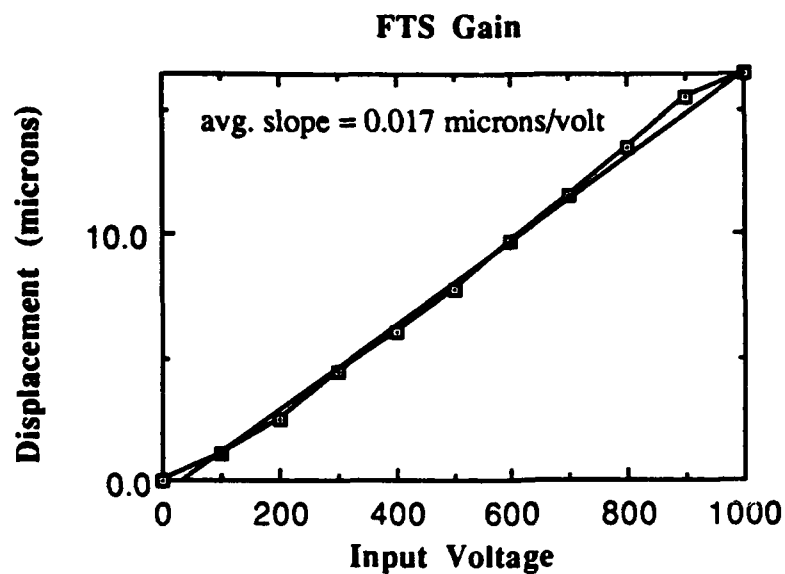
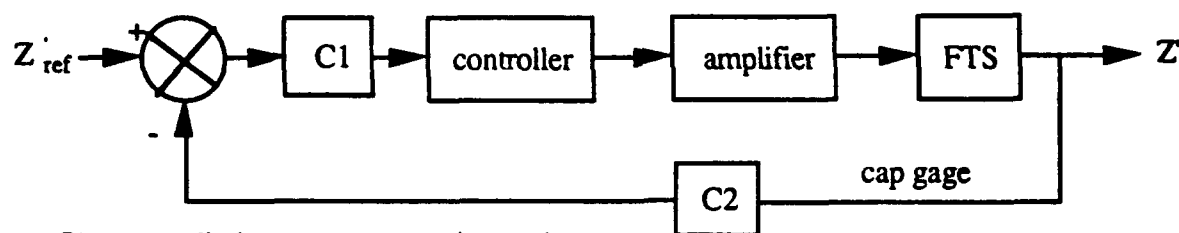
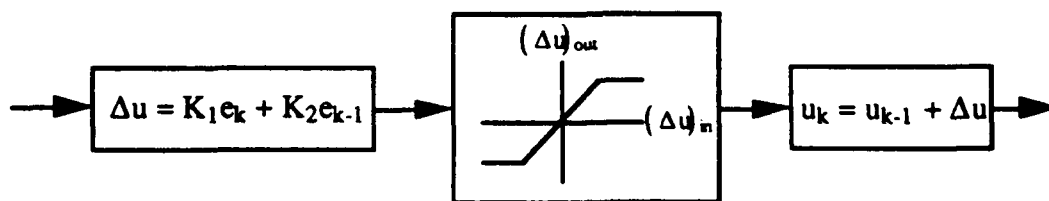


Figure 10: FTS displacement for static input voltage



C1 converts displacement error to an input voltage
 C2 converts cap gage voltage to displacement

(a) Block diagram of FTS system and controller



(b) Elements of the "controller"

Figure 11: Control system block diagrams

The servo and control system were tested by commanding a triangle wave of $1.5 \mu\text{m}$ amplitude. Figure 12(a) shows the open loop response. Although the response is very smooth, it doesn't achieve the commanded input amplitude. Figure 12(b) shows the closed loop response. Although not as smooth, its amplitude is much closer to the commanded amount. The control change or effort limit acts as a low pass filter because it prevents abrupt or high frequency changes to the control signal. Tightening the effort limit would smooth the closed loop response but also slow the response time.

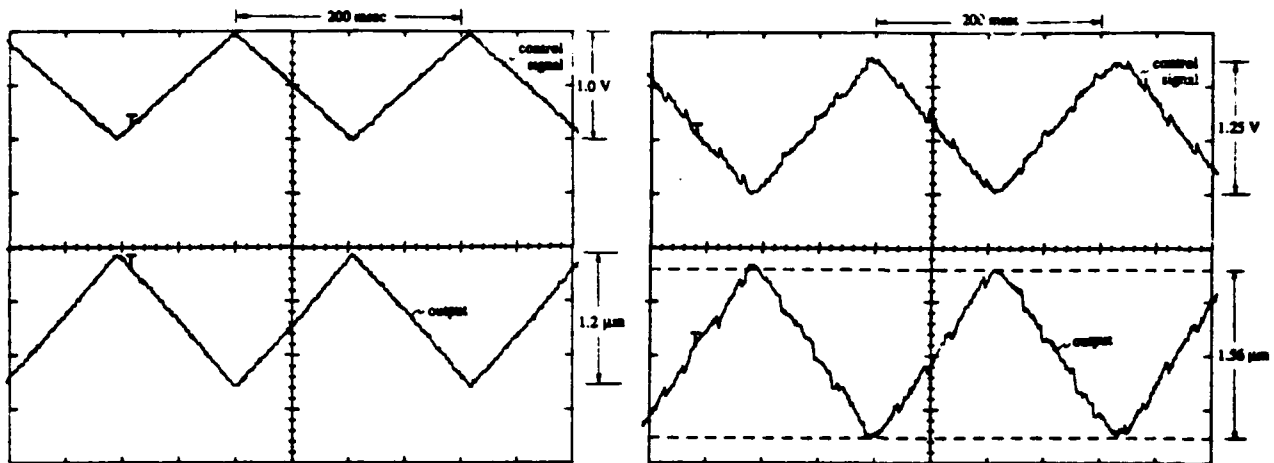


Figure 12: (a) Open loop and (b) closed loop responses of the servo system

An off-axis parabolic section was machined using closed loop servo control. Figure 13 shows an interferogram of the surface measured in its off-axis position [4]. While machining this surface, the servo cycled twice per part revolution (as shown in Figure 7). The interferogram has four lobes which indicate that the servo didn't exactly follow the reference signal. The amplitude of the servo motion was slightly less than intended. This error is probably due to the slow response time of the piezoelectric amplifier (Burleigh PZ70) used for this cut. Using an amplifier which supplies higher current would improve the response time.

16.4 CONCLUSIONS AND FUTURE WORK

The controller architecture which integrates DSP based FTS control with the 68020 based X and Z axis control has allowed implementation of slideway error correction and fast tool servo control algorithms. Ultimately the DTM controller will use fast DSP's for all of its control processes. With the H²ART architecture, which incorporates two DSP's (thus eliminating the 68020 processor), computationally intensive tool radius corrections could be implemented. In addition, a H²ART controller would have the capability to transfer previously calculated FTS reference points

from storage to the FTS control DSP. Such capability would reduce the amount of real-time computation and allow more complex FTS reference generation equations.



Figure 13: Interferogram of off-axis parabolic surface

References

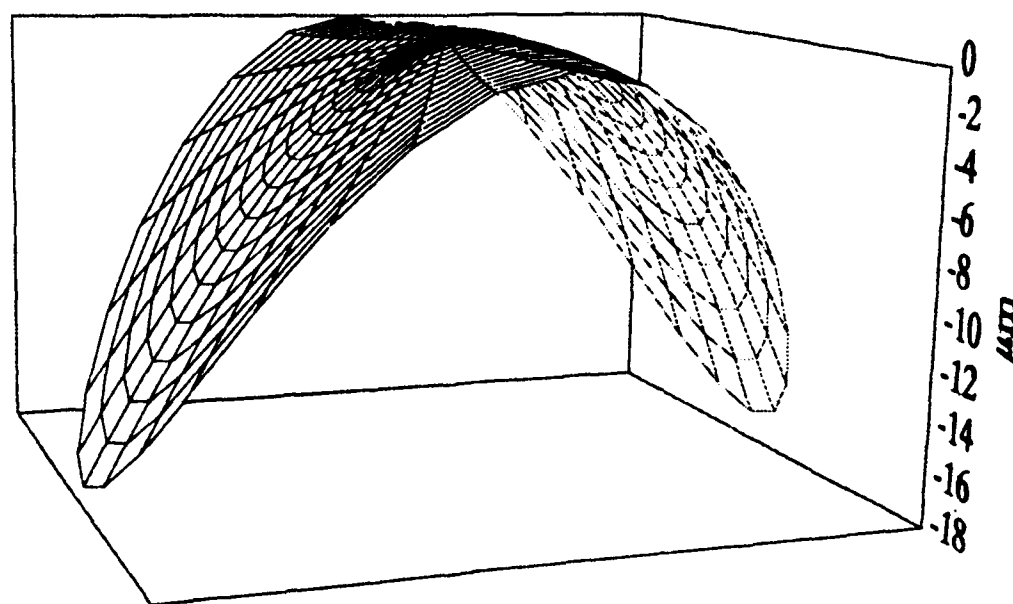
- [1] Gerchman, M. C., "An Exact Description of Far Off-Axis Conic Surfaces for Non-Rotationally Symmetric Surface Generation," *Proceedings of the ASPE Annual Meeting*, 1990.
- [2] Miller, M. H., Dow, T. A., "Controller Algorithms for Three Axis Diamond Turning Machine," *Precision Engineering Center 1989 Annual Report*.
- [3] Falter, P. J., "Diamond Turning of Nonrotationally Symmetric Surfaces," *PhD Dissertation*, North Carolina State University, 1990.
- [4] Maxey, L. C., "Novel technique for aligning paraboloids," Oak Ridge National Laboratory, 1991.

17 DIAMOND TURNING MACHINE CONTROLLER SOFTWARE DEVELOPMENT

Kenneth P. Garrard
Research Assistant
Precision Engineering Center
Robert J. Fornaro
Professor
Computer Science Department

The Precision Engineering Center's Fast Tool Servo has been successfully integrated with a Rank Taylor Hobson ASG-2500 Diamond Turning Machine and its PC-AT / 68020 controller. The DTM control computer also serves as the host for a single node H²ART system which controls the FTS actuator. Hardware interfaces have been constructed to provide both controllers with real-time access to critical machine data. Operation of the FTS is transparent to the DTM and its controller.

Non-Symmetric Z Axis Perturbation



17.1 INTRODUCTION

A controller for the Precision Engineering Center's ASG-2500 Diamond Turning Machine (DTM) has been developed over the past three years using a PC-AT / 68020 computer system [1,2]. This controller has provided improved performance and increased flexibility over commercial DTM controllers. However, it possesses neither the computational power nor the data transfer capabilities necessary for the implementation of an integrated Fast Tool Servo (FTS) controller. The H²ART architecture provides the processing speed, direct I/O interfaces, and interprocessor communication facilities required for this application [3]. An updated description of the H²ART architecture can be found in Section 18. The 68020 based controller remains useful as a testbed for algorithm development as well as rotationally symmetric diamond turning.

Previous FTS / DTM controller integration plans required the redesign of the axes control software for operation on a H²ART digital signal processor (DSP) [1,2]. A significant portion of the implementation effort for this design would be invested in replacing the existing axes controller. As an alternative, a new H²ART based FTS controller has been constructed that incorporates the existing 68020 DTM axes controller.

17.2 68020 CONTROLLER ENHANCEMENTS

Three principle enhancements to the 68020 based controller have been made. A new axes data collection program was written to aid in control algorithm development and vibration studies. Laser interferometer data can now be sampled and archived without enabling feedback control of the axes. The servo amplifier for each axis is enabled independently, thus preventing or allowing axis drift as desired by the user. As with all previous archive facilities, sample rate, averaging and delay parameters are adjustable. Two auxiliary analog input channels (e.g., capacitance gage, temperature sensor, LVDT) may be sampled in addition to interferometer data. This provides a time history of controller and machine activity as measured by both internal and external probes. An eighth-order laser position feedback filter has also been implemented. The purpose of the filter is to allow control in the feedback loop as well as in the feed forward loop. Evaluation of this control technique is in progress.

Geometric slideway error compensation has been implemented following the procedure outlined in [4,5]. The resulting improvements in controller performance are discussed in Section 16. Straightness and squareness error data for the p and z axes are obtained during the controller update cycle via table look-up. The z axis yaw error is calculated using a linear approximation. The slope of this line is fixed, but the intercept can be changed by the user. The error table for each axis is implemented as an Nx2 array, where N is less than 250. Each row of an error table contains a cumulative axis error value and the position of the other axis at which that error is

applicable. Linear searches through the error tables are avoided by noting that on each controller cycle the array index of the correct cumulative error is at most one table entry away from its position on the previous cycle. The correct choice depends on the velocity and direction of axes motion. The velocity of the p and z axes cannot be great enough to move through two error table entries during a single 1 KHz update cycle. Thus, the geometric error compensation procedure is independent of the size of the error tables and can be calculated in constant time.

17.3 H²ART SOFTWARE DEVELOPMENT

The supervisory software needed for H²ART program development and debugging has been completed. This software (*hh*) addresses the issues of node configuration (i.e., number and types of processors), program loading and startup, non-time critical communications, and host access to all Multibus and local bus memory resources. The host software consists of a function library and a user interface program. The library is structured so that applications programs (e.g., the FTS controller) can make use of its functions directly (i.e., without the user interface). The user interface program includes batch file and macro capabilities as well as command editing and history keystroke functions. A general purpose on-line help facility has also been implemented.

To provide host access to the local bus memory resources, a monitor program was written for the 80186 node control processor. This program responds to requests from the host sent via dual port memory resident on the BiT3 bus adaptor. Current capabilities include loading, dumping, and bitwise memory operations. Host command functions such as starting and stopping DSPs are formulated as sequences of these fundamental operations.

Performance of the H²ART multiprocessor system is critically dependent on the speed of the node controller. An efficient synchronous symmetric message passing algorithm has been designed to support both inter and intra node DSP communications on the H²ART system. A polling loop is used by each node controller to service requests from DSPs (and the host). The object of each poll loop test is always in the node controller's on-board memory. A pending request is sent from one node controller to another by modifying the poll object reserved in the memory of the other node for the specific communication action desired. In this way bus contention is minimized until the actual data transfer is initiated. Mutually exclusive access to shared data structures is not required. Future expansion of the node control monitor program will permit alternative real-time communications (e.g., asynchronous message passing, ADA rendezvous) and data buffering schemes to be evaluated for use by DTM, FTS, and other applications.

17.4 INTEGRATED FAST TOOL SERVO CONTROLLER

As an intermediate step in the process of developing a fully functional DTM capable of fabricating non-rotationally symmetric surfaces, an integrated (but add on) FTS controller has been built for application to current DTMs. In particular, an FTS and its H²ART controller have been interfaced to the PEC 68020 based controller and ASG-2500 DTM. The technique used is generic with respect other controller systems and DTMs [5].

Non-rotationally symmetric surfaces can be machined on-axis by using a high-bandwidth FTS actuator to generate small perturbation z-axis motions that are synchronized with the rotation of the workpiece spindle (θ) and the motion of the radial axis (ρ). The remainder (or base) surface is described by a best-fit aspheric surface of revolution. Being independent of θ , this base surface can be machined by a standard two axis DTM. The discussion below outlines the design of the integrated DTM / FTS system, the hardware interfaces needed to satisfy its intercontroller data flow requirements, and the derivation of the reference surface generation algorithm for off-axis conics.

The principle components of the system and their interconnections are shown in Figure 1. The FTS is driven by a high-voltage amplifier directly from the H²ART Analog I/O board. Servo position feedback via a capacitance gage also interfaced through this analog board. The FTS feedback control loop is essentially independent of the DTM slide control. However, generation of the non-symmetric reference signal requires access to the spindle encoder (θ) and the Zygo interferometer (ρ and z). Three logical interfaces are needed to build an integrated FTS controller that combines the H²ART, FTS Actuator, ASG-2500 DTM, Zygo Laser Interferometer, and 68020 axes controller components. These three interfaces are briefly described below. Further details can be found in Section 18.

17.4.1 Logical Interfaces

Angular Feedback Interface The Angular Feedback Interface provides the FTS controller with the current value of θ expressed as the number of encoder pulses since the last once-per-revolution index pulse. The maximum resolution of the encoder is 0.0015 radians per count ($\sim 0.1^\circ$). The direction of rotation can also be determined. The once-per-revolution index pulse is available externally to facilitate angular alignment of the workpiece fixture.

Laser Input Interface The Laser Input Interface maps the bits of the 50-pin Zygo Servo Interface connectors (one per axis) into one 8-bit control input port (both axes), one 8-bit control output port (both axes) and two 16-bit data input ports (one per axis). This gives the FTS controller timely access to the current ρ and z axes positions.

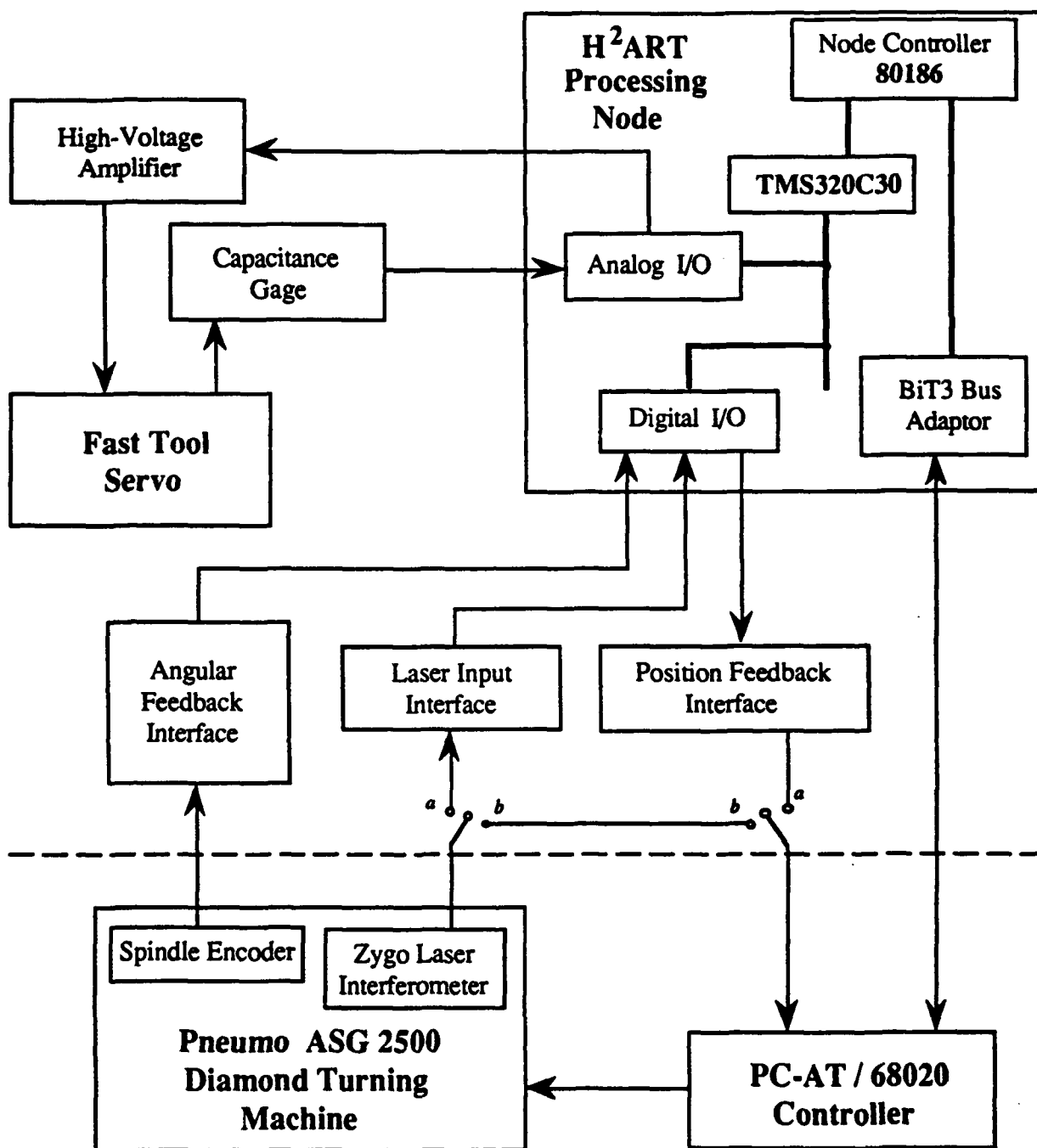


Figure 1. Integrated FTS Controller Configuration.

Position Feedback Interface The 68020 controller also needs access to ρ and z . However, it is not feasible to connect two controllers directly to the Zygo Servo Interface output ports, since they would both attempt to drive the measurement board sampling signals. The Position Feedback Interface replicates the signals that the 68020 controller expects during its communications cycle with the Zygo Interferometer measurement boards. It also provides one level of data buffering. This permits the FTS controller to communicate current axis position information to the 68020 controller without synchronization delays. The only requirement is that the FTS controller supply current ρ and z values to the interface board at a higher frequency than the 68020 controller update cycle (i.e., greater than 1 KHz). This is depicted in Figure 1 as switch position *a*. The Zygo Interferometer can be manually reconnected to the 68020 controller (switch position *b*) when the FTS is not in use.

17.4.2 Reference Surface Generation

The decomposition of the general aspheric optics equation into (ρ) -dependent and (ρ, θ) -dependent components suitable for on-axis FTS machining of off-axis conic segments has been described [6,7]. This solution minimizes the range of motion required of the FTS by tilting the workpiece. The best fit asphere is generated off-line as a sequence of cords. The number of cords can be adjusted to obtain any accuracy required in the asphere approximation. A high-speed axes control processor (i.e., H²ART node with floating-point TMS320C30) could also be used to generate the asphere *on-the-fly*.

The derivation of the surface equations and the method of dividing the total surface into symmetric and non-symmetric components is discussed below. The general equation of an optical surface of revolution about the z axis can be written as [8]:

$$z = \frac{cs^2}{1 + \sqrt{1 - (k+1)c^2s^2}} + a_1s^4 + a_2s^6 + a_3s^8 + a_4s^{10} \quad (1)$$

where

$$s^2 = x^2 + y^2 \text{ and } c = \frac{1}{r} = \frac{1}{\text{radius of curvature}}$$

The aspheric deformation constants (a_i) will be assumed to be zero so that the surface is a conic surface of revolution. The constant k is related to the eccentricity (e) of the conic ($k = -e^2$) and defines its shape as follows:

Surface	k
Hyperboloid	$k < -1$
Paraboloid	$k = -1$
Ellipse (rotated about major axis)	$-1 < k < 0$
Sphere	$k = 0$
Ellipse (rotated about minor axis)	$k > 0$

The geometry of the off-axis segment and its orientation in the parent coordinate system is shown in Figure 2. Fabrication of a segment of this surface by on-axis machining requires three coordinate transformations (translation, rotation and transformation to a cylindrical system) to produce the equation for the surface as a function of the radius and the angle of rotation of the spindle. These transformations have been published by Gerchman [6] and the resulting equations for the on-axis fabrication of a general conic surface of revolution (Equation (1) with $a_i = 0$) are given below:

$$z = d_1 + d_2 \rho \cos(\theta) - \sqrt{d_3 + d_4 \rho \cos(\theta) + d_5 \rho^2 + d_6 \rho^2 \cos^2(\theta)} \quad (2)$$

where the constants d_1, \dots, d_6 are given by:

$$d_1 = \frac{x_0 \cos(\alpha) + r \cos(\alpha) - (k+1) z_0 \sin(\alpha)}{1 + k \cos^2(\alpha)} \quad (3)$$

$$d_2 = \frac{-k \cos(\alpha) \sin(\alpha)}{1 + k \cos^2(\alpha)} \quad (4)$$

$$d_3 = d_1^2 \quad (5)$$

$$d_4 = 2d_1 d_2 - \frac{2 [x_0 \cos(\alpha) - r \sin(\alpha) + (k+1) z_0 \sin(\alpha)]}{1 + k \cos^2(\alpha)} \quad (6)$$

$$d_5 = \frac{-1}{1 + k \cos^2(\alpha)} \quad (7)$$

$$d_6 = d_2^2 - \frac{k \sin^2(\alpha)}{1 + k \cos^2(\alpha)} \quad (8)$$

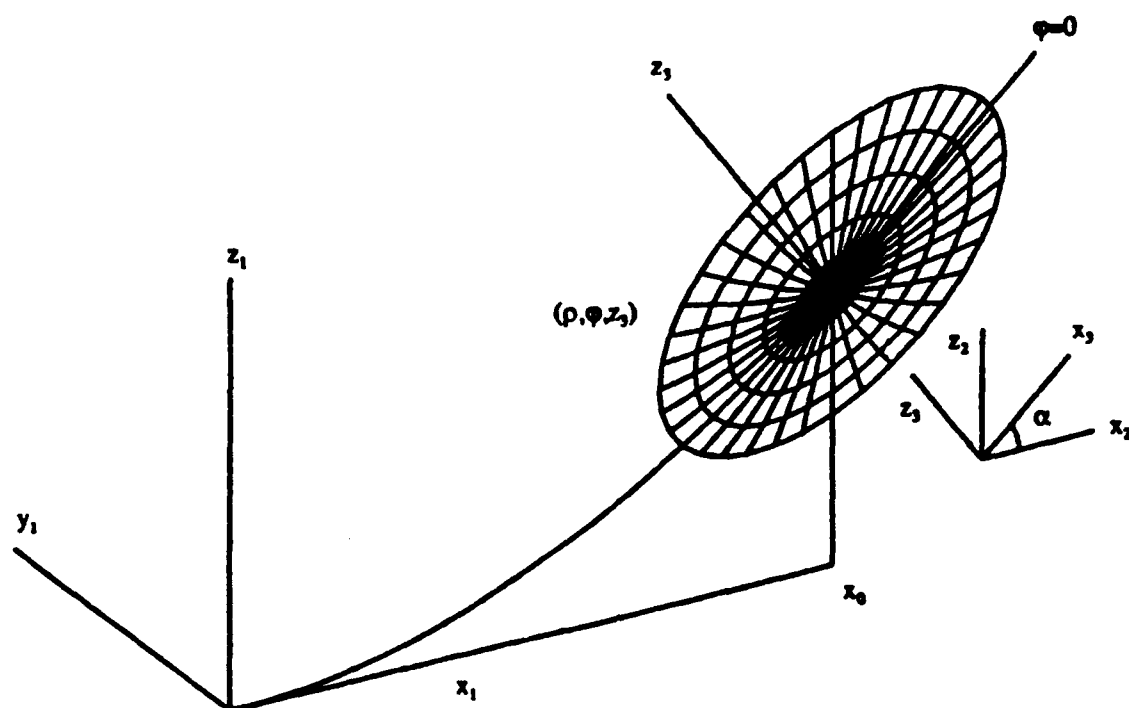


Figure 2. Geometry of Off-Axis Segment in Framework of Parent Conic [6].

Equation (2) gives the value of z as a function of the radius from the spindle axis (ρ) and the rotational angle (θ) of the spindle. It consists of a rotationally symmetric component (approximately 98% of the shape) and a non-rotationally symmetric component. The constants in Equation (2) which are defined by Equations (3-8) depend upon the conic constant, k , the off-axis translation parameters (x_0, z_0) and the angle of tilt (α) of the surface. The initial selection of the tilt angle is the tangent angle of the parent paraboloid at the center of the segment. However, this angle does not minimize the excursion of the fast tool servo. Tilt angle optimization is discussed in Section 16.

Once an acceptable value for α is determined the two components of the surface shape can be written as variations of Equation (2). The rotationally symmetric part will be the value of z at $\theta = \pi/2$, and the non-rotationally symmetric part will be the difference between the value of z from Equation (2) and the rotationally symmetric part.

The evaluation of Equation (2) using 32-bit floating-point arithmetic gives incorrect results due to the subtraction of two very large but nearly equal quantities. A further simplification is possible that solves this loss of precision problem. It also reduces the evaluation time of the function with a TMS320C30 by approximating the square root operation with a polynomial expansion. Using Equation (5), the value of z can be rewritten as:

$$z = d_1 + d_2 p \cos(\theta) - d_1 \sqrt{1 + \frac{d_4}{d_3} p \cos(\theta) + \frac{d_5}{d_3} p^2 + \frac{d_6}{d_3} p^2 \cos^2(\theta)} \quad (9)$$

The terms under the radical can be approximated using the following relation with $p = 2$:

$$(1 + x)^p = 1 + px + \frac{p(p-1)}{2!} x^2 + \frac{p(p-1)(p-2)}{3!} x^3 + \dots \quad (10)$$

The final simplified form of Equation (2) is now:

$$z \cong d_2 p \cos(\theta) - d_1 \left[\frac{1}{2} E - \frac{1}{8} E^2 + \frac{1}{16} E^3 - \frac{5}{128} E^4 \right] \quad (11)$$

where E is given by:

$$E = \frac{d_4}{d_3} p \cos(\theta) + \frac{d_5}{d_3} p^2 + \frac{d_6}{d_3} p^2 \cos^2(\theta) \quad (12)$$

Using Equation (11) the integrated FTS controller calculates the z -axis reference position (as a function of the current p and θ) in real-time and subtracts the current z axis position to obtain the small perturbation component. Using the actual axes positions in the reference calculation allows the FTS controller to compensate for high-frequency z axis vibrations as well as the low-frequency positioning error of the slower 68020 axes controller. There are several additional advantages to this approach. First, using existing hardware and software results in substantial savings in development time and effort. This technique is also portable to other DTMs and controllers.

17.4.3 Demonstration Part Generation

As a demonstration of the capabilities of the integrated DTM / FTS system, a 127 mm diameter segment of an off-axis parabola was machined. The radius of curvature of this parabolic segment was 2.159 m. Its location in the parent coordinate system was approximately 300 mm off-axis. A tilt angle (α) of 0.1364 radians resulted in a total z axis motion of 1.016 mm. Using Equation (11) the rotationally and non-rotationally symmetric components were calculated off-line. Plots of these two surfaces are shown Figures 3 and 4. Figure 3 shows the rotationally symmetric asphere which was machined by the DTM axes controller. The non-rotationally symmetric component machined on top of this asphere by the FTS is shown in Figure 4. The total range of motion required of the Fast Tool Servo was 17.78 μm or approximately 1.75% of the sag of the aspheric surface.

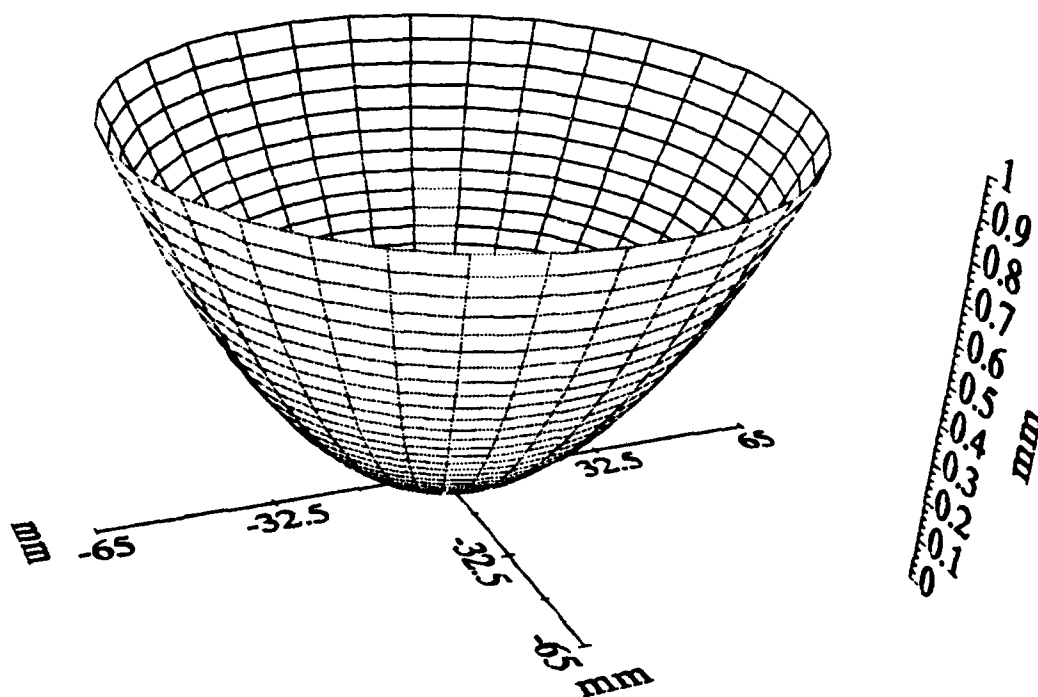


Figure 3. Rotationally Symmetric Aspheric Surface Generated by the DTM Axes Controller.

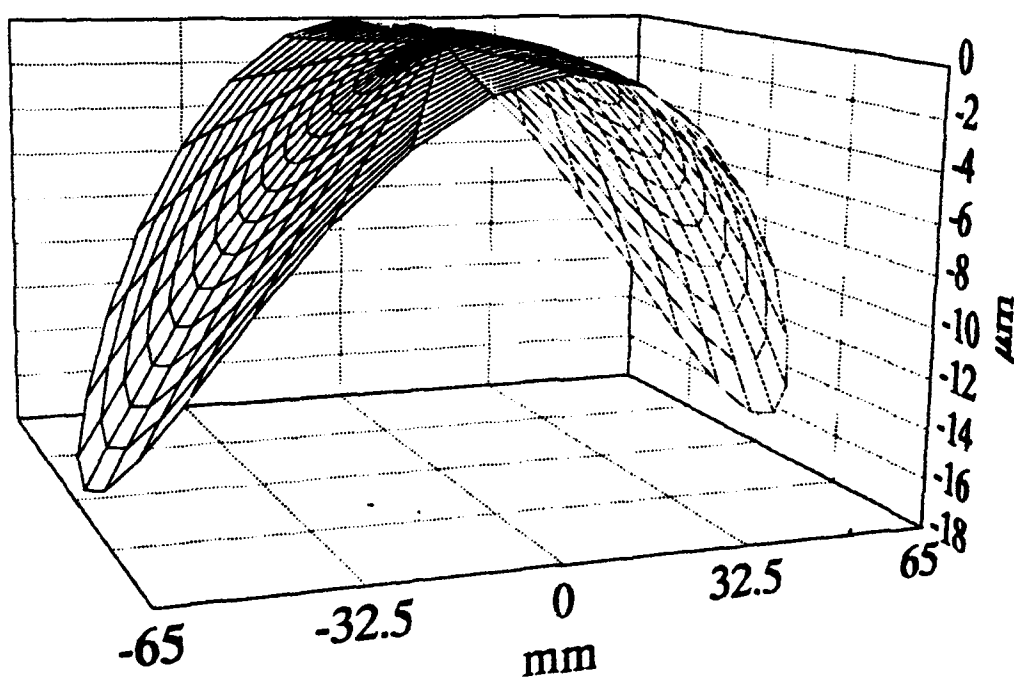


Figure 4. Non-Rotationally Symmetric Surface Generated by the Fast Tool Servo.

17.5 CONCLUSIONS

The 68020 based controller continues to be useful for DTM operations and algorithm development. Geometric error correction has been successfully implemented and should easily port to the new integrated FTS system. The H²ART supervisory software essential to applications development has also been completed. Real-time synchronization and communication issues can now be investigated using the H²ART platform.

Integration of the FTS system with the existing 68020 based controller and DTM components has been achieved. A H²ART system has been assembled for dedication to this project. Generalization of the FTS reference generation algorithm to incorporate the general optical surface equation is being studied. Tool path correction algorithms also need to be incorporated into both the DTM and FTS controllers. Though integrated at the hardware and computer systems level, much work remains before the user interface for the DTM controller is fully capable with respect to the FTS.

References

- [1] Garrard, K.P. and R.J. Fornaro, "Hardware/Software for Diamond Turning Machine Controller", *Precision Engineering Center Annual Report*, North Carolina State University, Raleigh, N.C., Vol. VII, pp. 27-36, December 1989.
- [2] Garrard, K.P. and R.J. Fornaro, "Diamond Turning Machine Controller Software Development", *Precision Engineering Center Annual Report*, North Carolina State University, Raleigh, N.C., Vol. VIII, pp. 307-318, December 1990.
- [3] Taylor, L.W. and R.J. Fornaro, "Development of Prototype Multiprocessor Architecture", *Precision Engineering Center Annual Report*, North Carolina State University, Raleigh, N.C., Vol. VIII, pp. 319-324, December 1990.
- [4] Miller, M.H. and T.A. Dow, "Controller Design for Three Axis DTM", *Precision Engineering Center Annual Report*, North Carolina State University, Raleigh, N.C., Vol. VIII, pp. 263-284, December 1990.
- [5] Miller, M.H., K.P. Garrard, T.A. Dow and L.W. Taylor, "Controller Design for a Modern Diamond Turning Machine", *Proceedings from ASPE Annual Conference*, pp. 62-65, October 1991.
- [6] Gerchman, M.C., "A Description of Off-Axis Conic Surfaces for Non-Axisymmetric Generation", *SPIE Proceedings*, Vol. 1266, 1990.
- [7] Luttrell, D.E., "Machining Non-Axisymmetric Optics", *Proceedings from ASPE Annual Conference*, pp. 31-34, September 1990.
- [8] Malacara Daniel, *Optical Shop Testing*, John Wiley and Sons, New York, 1978.

18 A H²ART BASED INTEGRATED FAST TOOL SERVO CONTROLLER

Lauren W. Taylor

Research Assistant

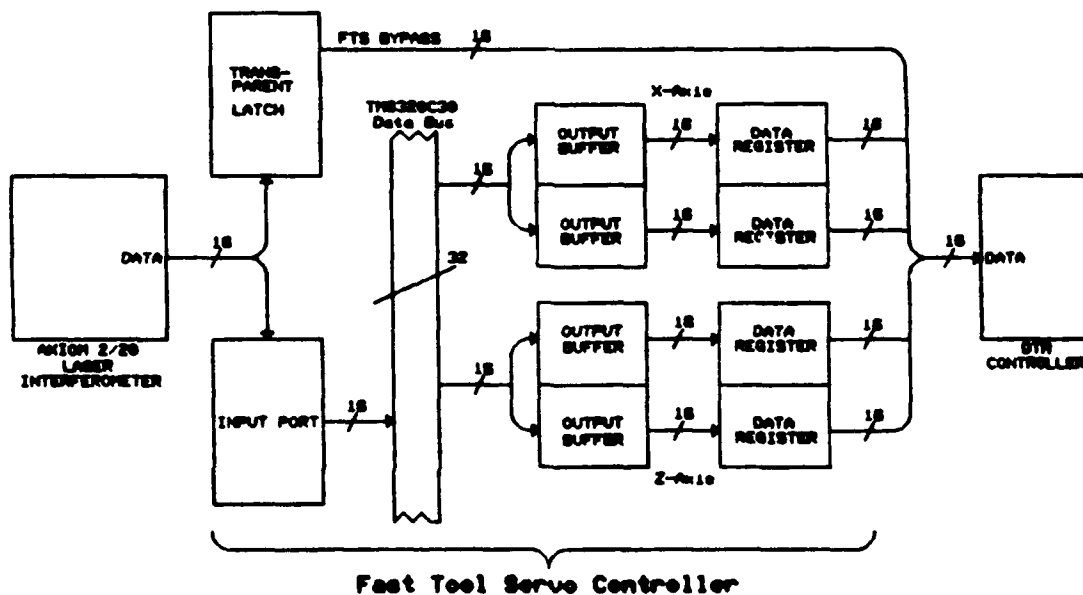
Computer Science Department

Robert J. Fornaro

Professor

Computer Science Department

A Heterogeneous Hierarchical Architecture for Real-Time (H²ART) has been implemented with the Texas Instruments TMS320C30 digital signal processor (DSP) by the Precision Engineering Center (PEC). While the ultimate goal is to integrate this architecture into the controller of PEC's diamond turning machine (DTM), an intermediate step was to add a fast tool servo (FTS) controller to the current DTM controller using the C30 based H²ART. A prototype FTS controller was built and successfully tested on PEC's diamond turning machine. This design is now being modified for use with other DTM controllers.



18.1 INTRODUCTION & BACKGROUND

The computer architecture being developed has been described as a Heterogeneous Hierarchical Architecture for Real-Time (H²ART). This description refers to a class of architectures which feature multiple heterogeneous processors connected through a hierarchy of busses to produce computational results in bounded, predictable time. Several examples of this architecture have been constructed at the Precision Engineering Center for use in advanced machine tool controller applications.

A H²ART system using an Intel 80186 node controller and Texas Instruments TMS320C25 digital signal processors (DSPs) was constructed at the Precision Engineering Center [1], but the lack of hardware floating point capability on the TMS320C25 limited its usefulness. The next generation of Texas Instruments' DSPs, the TMS320C30, had hardware floating point capability integrated into the processor and was found to have the computational speed necessary for real-time controller development.

18.1.1 The TMS320C30 Digital Signal Processor

The TMS320C30 is a 32-bit floating point digital signal processor with a minimum instruction cycle length of 60 ns. It has a 16M word external address space, 2K words of on-chip memory and a 64 word instruction cache. On-chip peripheral functions include two independent serial I/O ports, two counter/timers and a direct memory access (DMA) controller. The TMS320C30's instruction set provides for both integer and floating point arithmetic operations as well as interlocked operations for use in a multiprocessor environment. The instruction set also has provisions for parallel instruction execution. Texas Instruments software support for the TMS320C30 includes a macro assembler, linker, simulator and a C programming language compiler. An Ada programming language compiler is also available.

18.1.2 TMS320C30 Based H²ART System

A prototype TMS320C30 H²ART digital signal processor board has been constructed to evaluate it as a direct replacement for the TMS320C25 based version [1]. After the design was verified, two more prototypes were constructed on Multibus I (IEEE 796) form factor circuit boards to fit in the card cage with the node controller and I/O boards. These boards contain the TMS320C30, 128K bytes/32K words of memory shared with the node controller, 16K bytes/4K words of memory dual ported with the node controller, an input/output (I/O) interface and a serial interface between the DSPs. A block diagram of the H²ART system is shown in Figure 1.

The address, data, and control signals are interconnected between the node controller and DSPs with two 50-conductor ribbon cables. A third 50-pin connector on the DSP boards is the I/O interface. The serial ports are also connected by ribbon cable between the DSP boards. The connection to the Multibus I backplane is for power only.

Communications between the node controller and DSP are handled by the control and status ports. These ports also allow the node controller to start and stop the DSPs and determine whether the DSP or node controller has access to the 128K byte/32K word shared memory. Access to the dual port memory is arbitrated by hardware integrated into the dual-port memory chips.

The DSPs communicate with each other by either requesting the node controller to transfer data or directly through a serial interface. The serial ports that are integrated into each DSP can transfer 8, 16, 24 or 32-bit words. The transfer rate is programmable with a maximum speed of approximately 130K 32-bit words per second. The serial ports are doubled buffered and memory-mapped. After the initial programming, they require very little software overhead for operation (1 memory write and 1 memory read per transfer). The serial ports also provide "transmitter buffer empty" and "receiver buffer full" interrupts to both the central processing unit (CPU) and the DMA controller for interrupt driven applications.

18.2 FAST TOOL SERVO CONTROLLER

The fast tool servo (FTS) controller uses a H²ART system with a single DSP node as illustrated in Figure 2. The DTM controller serves as host for the H²ART system. The FTS controller requires data from three sources to compute the correct servo position: from the laser interferometer to determine the position of the slides, from the shaft encoder on the DTM's spindle to determine the angular position of the spindle, and from the capacitance gage mounted in the servo to determine the position of the tool. The output from the laser interferometer and shaft encoder is digital while the capacitance gage data is analog.

The FTS controller has output requirements in three areas: digital signals to control the operation of the laser interferometer, an analog signal to control the servo position; and digital position data that must be transferred from the interferometer to buffers which can be accessed by the main controller. The last item is necessary because both the FTS controller and the main controller need position data from the laser interferometer to function properly and it is not feasible to let both controllers attempt to drive the laser control signals.

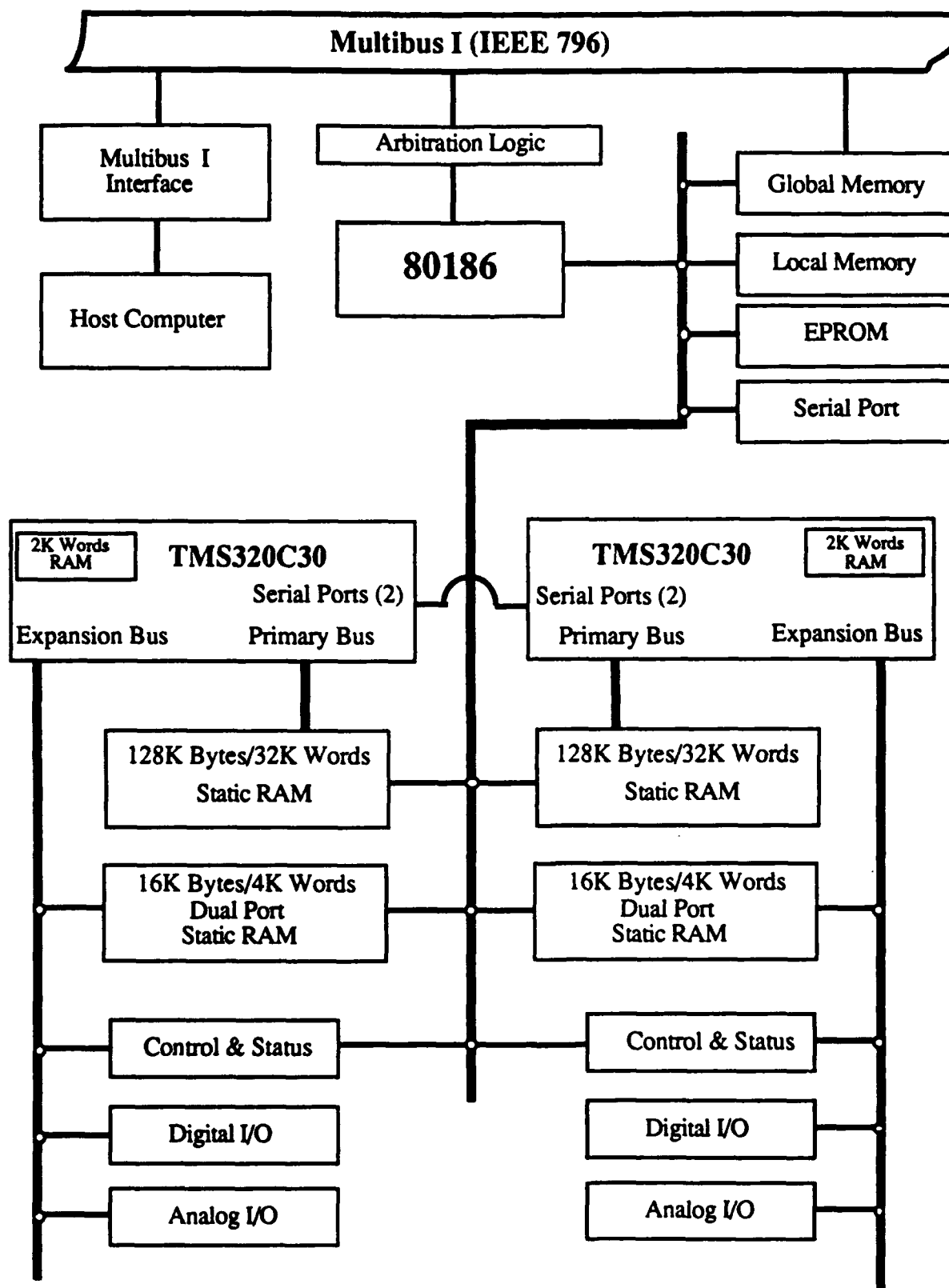


Figure 1. H²ART node with two TMS320C30 DSPs.

The H²ART system has four interfaces, shown in Figure 2, to meet the FTS controller's requirements: a laser input interface that controls the laser interferometer's operation and reads the slide position data, a position feedback interface that gives the main controller access to the laser data when it is needed, an angular feedback interface that gives the FTS controller the spindle's angular position, and an analog input/output (I/O) interface that moves the FTS and determines its position. The first three interfaces (which are all digital) are constructed on a single Multibus I (IEEE 796) form factor board, and a second Multibus I board has the analog I/O interface. The Multibus I form factor is used for convenience of packaging, and the only connections from these boards to the Multibus I backplane are for power. A detailed discussion of these interfaces follows.

18.2.1 Laser Input Interface

The Axiom 2/20 laser interferometer used in PEC's diamond turning machine is an older version with model 7064 measurement boards that operates only in the singular mode. The measurement board of each axis is connected to the FTS controller (or the main controller if the FTS controller is not present) by its individual 50 conductor ribbon cable. Newer versions of this interferometer, with model 7064A measurement boards, can operate in the multiplexed mode where all measurement boards (up to 16) are connected to the controller through a single 50-conductor ribbon cable. The configuration of the laser input interface depends upon the interferometer's mode of operation. Control and status signals in the singular and multiplexed modes also differ, so the position feedback interface circuitry is also dependent on the interferometer's mode of operation.

In the singular mode of operation, the controller starts the interferometer read cycle by asserting the SAMPLE signal. The measurement board responds with a TRANSFER signal when the laser data is latched into its data registers. The controller then asserts the READ LOW WORD signal which places the 16 least significant bits (LSB) of the laser information on the data lines, then reads and stores the data. Next, the controller asserts the READ HIGH WORD signals and reads the 16 most significant bits (MSB) of data which completes the cycle. This whole cycle is repeated for each measurement board in the system.

The interferometer read cycle in the multiplexed mode also starts with the controller asserting the SAMPLE signal. Since the TRANSFER signal is not available in the multiplexed mode, the controller must wait a minimum of approximately 400 nS before attempting to read the laser data. In addition to asserting the READ LOW WORD and READ HIGH WORD signals, the controller must also furnish 4 bits of address information to assure that the correct measurement board responds to these signals and places its data on the bus. The controller reads data from subsequent measurement boards by asserting the READ HIGH WORD and READ LOW WORD signals with the address of the board that it wishes to access.

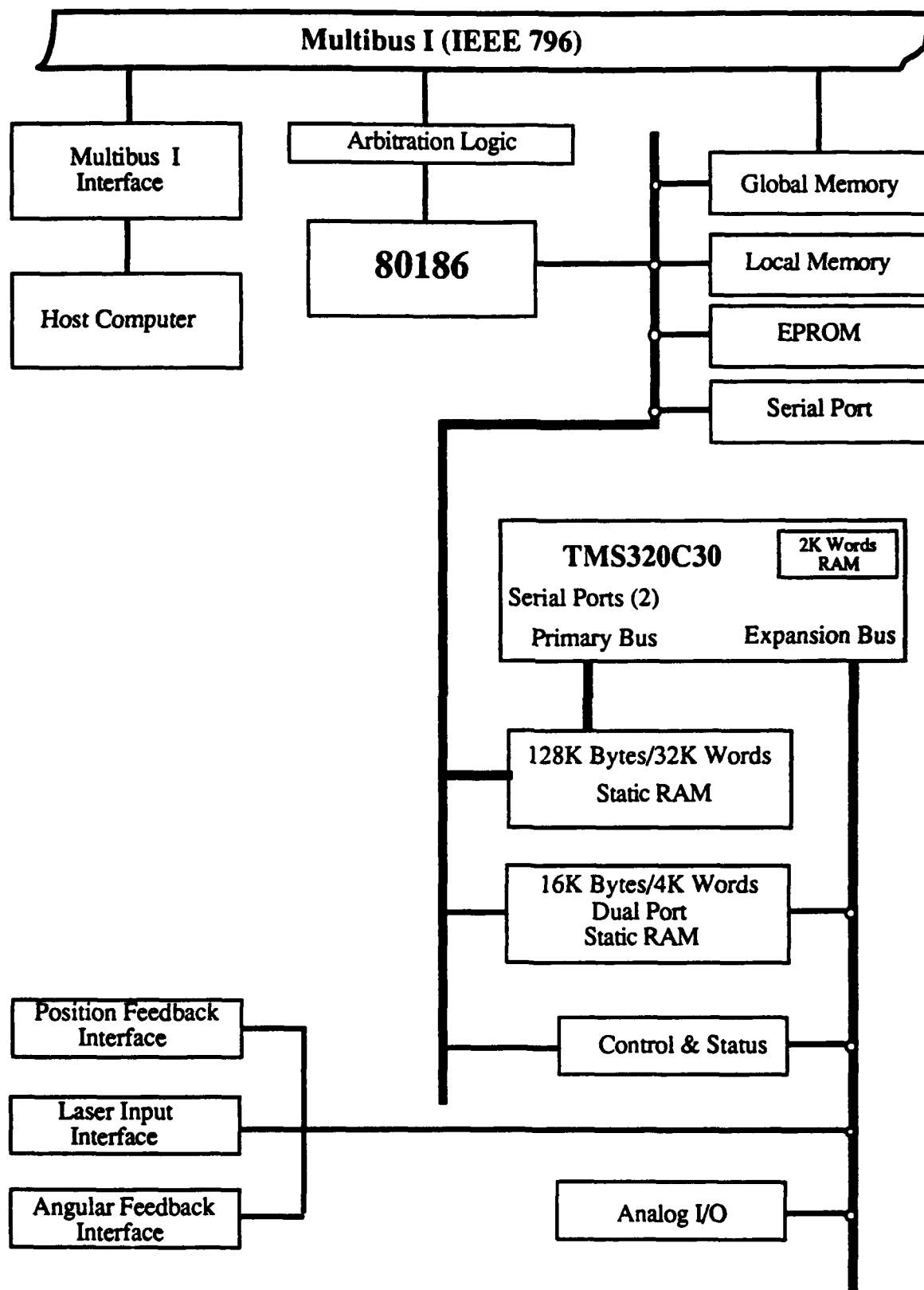


Figure 2. H²ART Node with FTS Controller Interface

In addition to the TRANSFER signal in the singular mode, the laser interferometer generates an ERROR signal in either mode of operation. This signal indicates an error condition in the laser interferometer system and is connected to both the FTS controller and the main controller.

A set of transparent latches which bypasses the FTS controller and connects the main controller directly to the laser interferometer is included in the design. This arrangement allows the FTS controller to be removed from the control loop without unplugging and reconnecting cables.

18.2.2 Position Feedback Interface

Both the main controller and the FTS controller need slide position information from the laser interferometer to operate properly. Since the FTS controller needs this data at a higher rate than the main controller, it controls the operation of the laser interferometer and writes the slide position information to a set of buffers as it reads it. The main controller's access to these buffers is through circuitry that emulates the laser's response to control signals. In other words, the main controller thinks it is communicating with a laser interferometer but it is actually communicating with an interface between it and the FTS controller. The main problem in the design of the position feedback interface was to insure that the main controller was not given access to the slide position data until the laser input interface had written both 16-bit halves of the data word to the output buffers.

The PEC's laser interferometer operates in the singular mode, and its main controller waits for the measurement board to assert its TRANSFER signal before it attempts to read the position data. The position feedback interface in the PEC's FTS controller delays this signal if the laser input interface has not written both halves of the data word to the output buffers. This scheme does not work if the measurement boards of the laser interferometer are operating in the multiplexed mode because the TRANSFER signal is not available. The data path through the FTS controller interfaced to an Axion 2/20 interferometer operating in the singular mode is shown in Figure 3.

The position feedback interface being constructed for use with multiplexed mode interferometers has a second set of buffers or data registers (Figure 4). The position data from the output buffers is latched into these registers when the main controller asserts the SAMPLE signal and it remains in these registers until the main controller needs more slide position information and asserts the SAMPLE signal again. If the laser input interface has written only one half of a data value to the output buffers, the transfer of data between the output buffers and the data registers is delayed a maximum of 375 nS.

In addition to reading the position data from the measurement boards the controller that is interfaced to the laser can set the accumulators on these boards to a known state (all zeros) by asserting the INITIALIZE signal. In the singular mode, the signal is sent to each measurement board through its individual 50-conductor ribbon cable, while in the multiplexed mode it is sent through a common connection with address information which routes the signal to the correct measurement board. In the PEC's FTS controller, the INITIALIZE signal goes directly from the main controller to the laser interferometer bypassing the FTS controller. In the controller being constructed for use with multiplexed systems, the INITIALIZE signal, along with the address data from the main controller, is decoded and used to inform the FTS controller that the main controller wants a measurement board to be initialized. The FTS controller's laser input interface then asserts the INITIALIZE signal along with the proper address information to service the main controller's request.

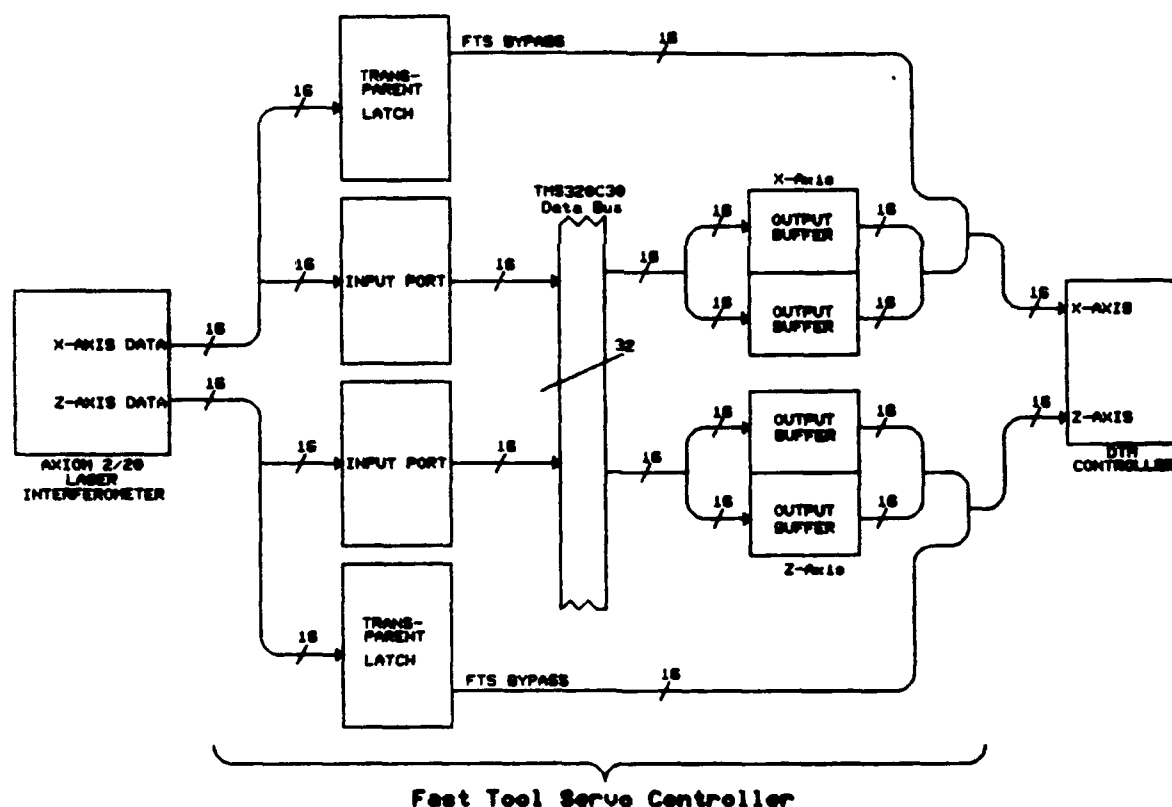


Figure 3. Data Path Through FTS Controller (Singular Mode)

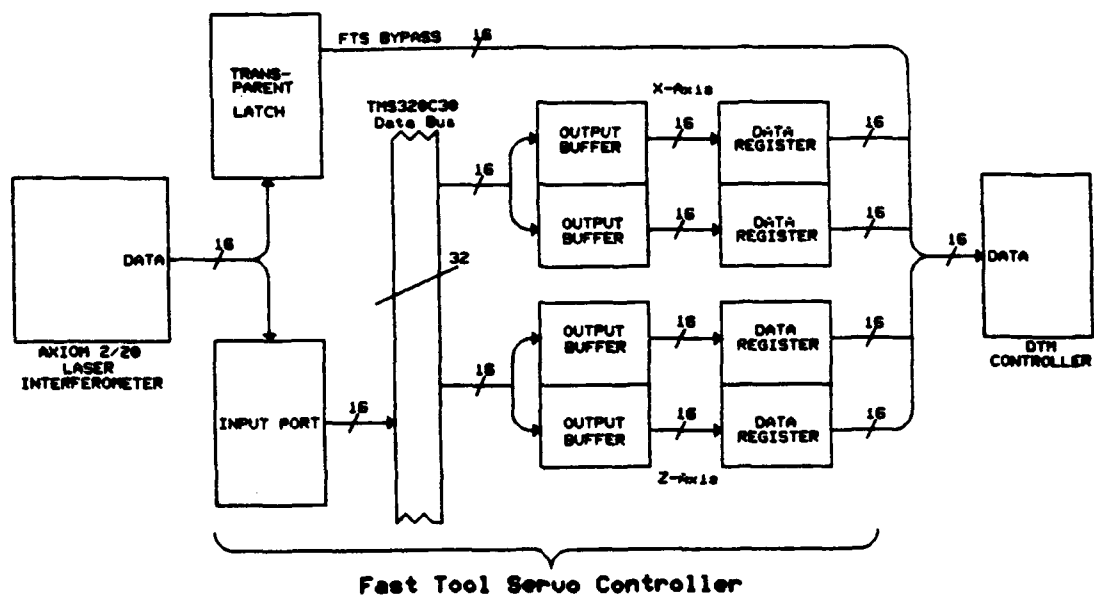


Figure 4. Data Path Through FTS Controller (Multiplexed Mode)

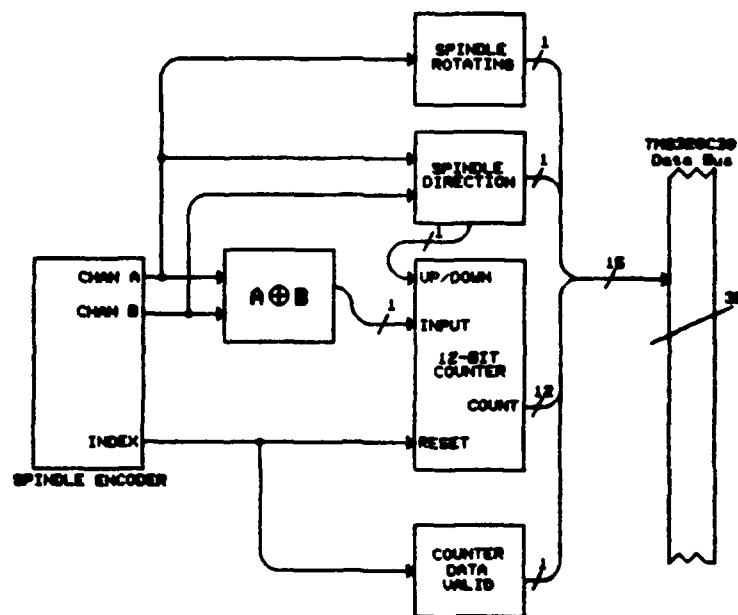


Figure 5. Angular Feedback Interface

18.2.3 Angular Feedback Interface

The output from the spindle encoder on the PEC's diamond turning machine consists of two of 2048 pulse-per-revolution channels in quadrature (*i.e.* the channels are 90 degrees out of phase) plus a once-per-revolution index pulse. The FTS controller's angular feedback interface (Figure 5) exclusive-ORs the two 2048 pulse-per-revolution channels to get a 4096 pulse-per-revolution pulse stream. Instead of using these pulses to interrupt the FTS controller as was done in the past with open-loop FTS controllers, the pulses increment or decrement a 12-bit counter depending upon the direction of the spindle rotation. The FTS controller can read the value of this counter and determine the angular position of the spindle. The spindle encoder interface also has bits to tell the FTS controller if the spindle is rotating, its direction of rotation and whether or not the data in the counter is valid. The once-per-revolution index pulse resets the counter at the end of each revolution and is also used to determine the angular alignment of the workpiece.

18.2.4 Analog I/O Interface

The analog I/O interface has four channels of analog output and three channels of analog input, but only one output and one input channel is used in this application. The analog output channel uses a 16-bit digital-to-analog (D/A) converter with a fixed range of -10 to +10 volts. The analog input channel uses a 12-bit analog to digital (A/D) converter whose range is programmable by jumpers on the board. In this application, the A/D is programmed for a -10 to +10 volt range. The A/D's status bit (which tells if a conversion has been completed) can be read by the H²ART through the data bus or it can be connected to a latch on the analog I/O board, which will interrupt the DSP when a conversion has been completed. The A/D conversion time is approximately 15 μ S.

18.3 FUTURE WORK

A FTS controller is now being constructed to interface with the Axiom 2/20 laser interferometer operating in the multiplexed mode. This controller being built to operate in conjunction with the IBH controller on the Rank Taylor Hobson Nano Form 600 DTM at Oak Ridge National Laboratory. This will allow the DTM to machine non-rotationally symmetric components.

References

1. Taylor, L.W. and Fornaro, R.J., "Development of Prototype Multiprocessor Architectures", *Precision Engineering Center 1990 Annual Report*, Vol VIII, 1991, pp.323-324.

19 PERFORMANCE OF INTERPROCESSOR COMMUNICATIONS ARCHITECTURES

William D. Allen

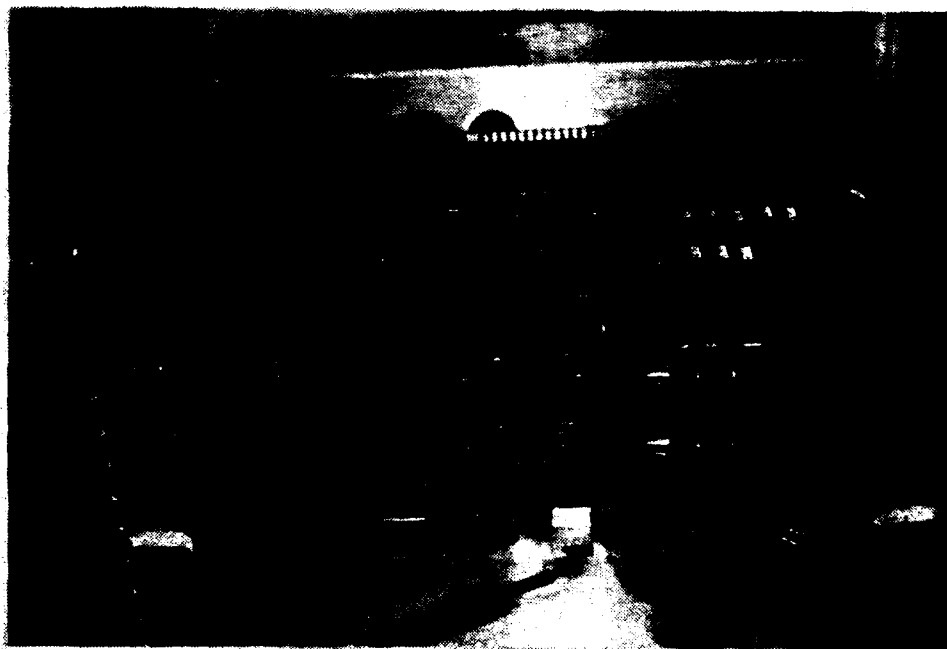
Graduate Student

Robert J. Fornaro

Professor

Department of Computer Science

Advanced control of high speed, precision processes requires high performance computing capability. The evolution of microprocessor technology has made it practical and cost effective to construct multiprocessor computer systems to satisfy this requirement. Such systems allow the control process to be subdivided into separate tasks which can be assigned to dedicated processors. Coordination of a control process requires that these processors communicate with each other. The ability to communicate efficiently is critical to the performance of a multiprocessor based control system. To design a system with reliable and predictable behavior, the impact of the communications architecture on the system must be understood.



19.1 INTRODUCTION

The most important property of a real-time computing system is its ability to perform needed computations within deadlines which must always be met. When the computational load exceeds the capacity of a single processor, an increasingly common solution is to divide the application into multiple parts for execution on a multiprocessor computer system.

Application of multiprocessor systems to real-time control can be divided into three research problems: decomposing applications into multiple tasks, assignment of tasks to processors, and development of efficient multiprocessor architectures to support these applications. The topics of application decomposition and task assignment have been previously addressed at the PEC [8]. Preliminary work by PEC researchers has been done in the area of determining the timing characteristics of systems given an application decomposition and task assignment [1], but extensions are required to make it generally useful. The ultimate objective is to provide methodology and tools to guide the development of real-time multiprocessor control systems with guaranteed performance characteristics. The integration of solutions to these problems remains an open problem in computer science research.

A thorough understanding of the characteristics of multiprocessor computations both from a software and a hardware standpoint is key to the development of these goals. The most significant hardware/software issue in the performance of a multiprocessor system is interprocessor communication. Since the tasks of a multiprocessor system are not autonomous, the time required to communicate with each other can be a significant factor in system performance. While the issues of communicating tasks have been addressed, many have made the assumption that tasks can communicate when they need to without delay. Issues such as the use of a common path for communication (which must be shared) and synchronization of tasks due to communications have not been adequately addressed. Research at the PEC into the impact of communications characteristics on system performance has two components. From a theoretical or analytical standpoint, the impact of communications and resource sharing on task schedules and deadlines has been studied and theories for establishing time constraints which can be guaranteed have been derived. The other component is an experimental study of the performance of different interprocessor communication architectures. From this study the effects on performance of various architectural elements in a multiprocessor system will be assessed. This will lead to recommendations for improved architectural characteristics of future multiprocessor systems for real-time control applications.

This section reports the results of continuing work relating to developing methodologies for designing real-time control systems with guaranteed performance characteristics. Four major topics are discussed. These are: 1) an analytical study of the impact of interprocessor communications over a common communications resource, 2) a survey of the evolution of

multiprocessor communication architectures, 3) a summary description of research multiprocessor systems being evaluated and 4) the results of experimental studies.

19.2 IMPACT OF COMMUNICATIONS PERFORMANCE

In any computing system, the speed at which the required computations can be accomplished is an important factor. Computational speed is the major element in any system model, but in a multiprocessor system, the effects of interprocessor communication must also be considered. Since data passage between processors is an important and potentially time consuming part of the application, accurate characterization of this process is necessary for developing an accurate model. Not only must the time required to pass data from processor to processor be evaluated but the impact on computational tasks due to communications must be recognized. Previous work [2] has studied the effects of blocking in a multiprocessor pipeline due to a shared communications resource. The impact of this blocking on a pipelined system's cyclic deadlines was studied and a method developed for establishing the minimum cycle time which can be guaranteed. Section 19.2.1 summarizes these results. The work on blocking in pipelined systems has been extended to cover the more general case of blocking due to the use of any type of shared resource by a non-pipelined system. This extension is reported below in section 19.2.2.

19.2.1 Pipelined Applications

Cyclic serial computational applications can be decomposed into multiple tasks for implementation on a multiprocessor system. Such a decomposition results in a pipelining of the computation with each processor being a stage of the pipeline. In some architectures, a shared resource (memory, bus) may be utilized for inter-task communication.

Pipelined processing has been widely used to enhance the performance of hardware systems. These same principles can be equally applied to a multiprocessor software implementation. The cycle time of the pipeline is set by the execution times of the stages and the means of communication (data passing) between stages. In a hardware system, each stage of the pipeline has the same execution time. However, a pipelined implementation of a software process may have different execution times for each of the stages. If the decomposition of the application is such that the computation is equally distributed, the optimum pipeline cycle time will result. In practice, many application decompositions tend to be coarse-grained resulting in an imbalance of task execution times. In a multiprocessor pipeline with unequal task execution times, the cycle time of the pipeline is set by the task with the longest execution time.

There are numerous mechanisms for interprocessor communications. This study has focused on a shared communications resource. This provides maximum interconnection flexibility but imposes the possibility of resource contention, thus use of the resource must be scheduled. Without some

scheduling regime in place, it is possible for one task to unduly delay another by seizing the common resource thus causing a deadline failure due to blocking. Since a resource, not a task, is being scheduled the hard real-time scheduling algorithms normally used for task scheduling do not directly apply. However, they provide guidance in the development of a resource scheduling protocol for the applications under consideration. Even with a resource scheduling regime in place, the problem of priority inversion caused by blocking delays can occur in multiprocessor systems just as in uniprocessor systems. The immediate objective is to define an efficient method for establishing and guaranteeing a cyclic schedule for pipelined computations.

In performing an analysis of the blocking patterns of a task set, the possibility of the occurrence of chains of blocking must be considered. That is to say that the blocking delay of one task could cause that task to create a blocking delay on another task. Likewise, the adjustment of the execution time of one task to eliminate a critical blocking could cause a blocking delay to appear elsewhere. Thus the blocking analysis/execution time adjustment must be an iterative process.

Solution of this blocking delay problem by structuring the system such that blocking does not occur is impractical. In theory, the execution time of fixed tasks can be adjusted so that contention does not occur. In practice, however, ensuring that all tasks have a fixed duration implies the necessity of analyzing the timing of all possible paths through the task code. Development tools to facilitate this analysis are not yet available. Even with such tools, retiming of programs would be computationally intractable. Further complicating the issue are timing and speed uncertainties of individual processors in a multiprocessor system, which could negate such designs.

Since the prevention of communication blocking delays is impractical, the solution to deadline assurance is to allow time within the computational cycle for blocking to occur. To accomplish this, a means to determine and bound the worst case blocking time must be established. Solution of the delay problem when tasks have variable execution times is simplified by the constrained nature of a pipelined particular application, primarily by the fact that all tasks have the same periodicity. As noted before this allows a static solution to the resource scheduling problem to be formulated. Thus, in this approach, delays are allowed to occur and the worst case delay of the critical task is bounded. The pipeline cycle will be established by the execution time of the critical task plus the worst case blocking delay caused by other tasks.

Researchers at Carnegie-Mellon University (CMU) developed priority inheritance scheduling protocols to address an analogous problem in uniprocessor systems [6]. Their work related to task blocking by critical section usage. These techniques limit the blocking of a task to the duration of one (the longest if they are not equal length) critical section and assures that deadlock will not occur. Research at the PEC utilizes the priority inheritance concept to develop a theory to address the communications blocking issue in multiprocessor systems.

These theoretical results provide a methodology for establishing a bound on the delay caused by waiting for a communications resource. This bound is then added to the worst case cycle time of the longest pipeline task giving the minimum pipeline cycle time which can be guaranteed. The maximum delay for access to a communications resource is that which can be caused by a single usage of the resource by a task of lower priority. Therefore, since a task in the pipeline utilizes the shared communication resource twice (receive data and send data) this blocking can occur two times. Hence the worst case blocking delay of the critical (longest) task in the pipeline will be the sum of the two longest usages of the communication resource (excluding those which are part of the critical task). Priority for access to the communications resource is set by assigning priorities according to the slack time of each pipeline task. Thus the highest priority will be assigned to the critical task (zero slack time) and the lowest to the task with the most slack time.

To demonstrate the validity of this theory, the four processor pipelined application described in the paper reporting this work [2] was implemented. The results of the experimental demonstration followed the predictions of the theory and are summarized in section 19.5.2 of this report.

19.2.2 Generalization of Theory

The theory on assuring cyclic deadlines was developed in the context of pipelined applications. Further study has shown that the results can be extended to any multiprocessor application where a shared resource is used. Many of the constraints imposed by the structure of a pipelined system (identical cycle time for all tasks, all tasks dependent on each other, etc.) are not necessary conditions for occurrence of deadline failure due to blocking. A multiprocessor system with groups of tasks which communicate within the groups but not between groups can experience the same failures. While synchronous interprocessor communications in a pipelined system is a major example of a shared resource, it is not the only possible instance. A data structure in global memory which is locked while in use by a processor possesses the same characteristics as synchronous communications, i.e. any other requests for the resource are blocked until the current operation is completed and the resource released.

As an example of this extended scope, consider the execution graph shown in Figure 1. This depicts the multiprocessor execution of two independent tasks with different cycle times which use the same shared resource.

```
P1: ---**-----**-----**-----**-----**-----**-----**-----**-----
P2: -----**-----**-----==**-----**-----==**-----**-----
```

[==: blocked waiting for resource, **: resource in use, ---: computation]

Figure 1: Shared Resource Utilization

Task 1 has a cycle time 6 units of time and task 2 has a cycle of 8. Both tasks use the resource for 2 units. As can be seen from this graph, the task in processor 2 will fail to meet its 8 time unit deadline every other cycle due to the resource being locked by processor 1's task. If, following the guidelines developed for pipelined systems, the tasks are given 2 units of idle time (slack) on each cycle they will then be able to guarantee their deadlines.

This extended study has determined that the theoretical results for pipelined processes apply as well to the general case of any multiprocessor application using shared resources. Therefore the general case can be stated as follows:

The maximum delay which can be imposed on a task waiting for access to a shared resource is that which can be caused by use of that resource by a task of lower priority. Thus the minimum cycle time which can be guaranteed is equal to the worst case execution time of the task plus the product of the worst case resource delay and the number of times each resource is used by that task.

19.3 MULTIPROCESSOR ARCHITECTURES

The algebraic result established above identifies software engineering principles that can be used to design real-time systems. Application of these principles to account for the impact of interprocessor communications provides an important element in designing systems with guaranteed cyclic deadlines. A key element of this design process is an understanding of the characteristics of interprocessor communications. These characteristics are greatly affected by the architecture of the multiprocessor system's communications capability.

There are numerous multiprocessor computer system architectures. These architectures are broadly categorized based on the means of data exchange between processors. This is commonly referred to as the coupling of the system. Two architectures which are of the most interest for real-time control applications are the *tightly-coupled* where processors communicate with each other through a common memory or directly over shared or dedicated buses and *closely-coupled* where processors communicate via high speed interprocessor buffers or links. A third architecture, *loosely-coupled*, is not of major interest for real-time control applications under consideration here.

19.3.1 Interprocessor Communications Architectures

A key factor in the performance of a multiprocessor system is the speed at which data may be exchanged between processors. As a result, ongoing research is focusing on the performance impact of different hardware architectures for interprocessor communication support. The mechanism for interprocessor communications falls into one of two general categories, shared memory access and message passing. In shared memory access, data shared between processors

is kept in a area of memory accessible to all processors using that data. Here the movement of data between processors is a passive process. In message passing, the data is actively transmitted between processors in the form of discrete messages. (It should be noted that message passing can also be implemented using shared memory.)

Interconnection paths between processors generally fall into one of two categories, dedicated and shared. In a system using dedicated paths, processors are connected by discrete paths with each path serving only two processors. This is the classic concept of pipelining. Interference (blocking delays) due to interprocessor communications is not a factor. Dedicated paths usually exhibit blocking send/receive characteristics.

When shared interprocessor paths are used, processors are connected by a communications media which is shared by all processors. Thus the possibility exists of one interprocessor transfer being blocked while waiting for another transfer to complete. Several mechanisms have been employed to implement such a communications system. Among them are:

1. Global Memory: Data is passed through a memory accessible by all processors.
2. Memory Mapping: Each processor sees part of the other processors' memory as part of its own address space.
3. Communications Networks: Data is passed across a common communications network. These networks can be LAN's such as Ethernet or Token Ring for a loosely coupled system and interprocessor buses for a tightly coupled system.
4. Data Mover: Movement of data between processors is handled by a separate processor dedicated to handling communications between processors.

19.3.2 Development of Multiprocessor Interconnections

Hardware support for interprocessor communications is commonly a tradeoff between the cost of providing data paths and the facility with which data may be exchanged. The ideal communications facility would be one in which any processor can pass data to any other processor. Hardware supports the two general interprocessor communications methodologies by providing mechanisms whereby multiple processors can access the same physical memory (shared memory) or by providing mechanisms for message transport between processors.

Two classic architectures which provide this capability are the fully connected and crossbar networks. In a fully connected network, every processor is connected to every other processor by with a bidirectional path. This approach rapidly becomes expensive since each processor in a n -processor system must support $n-1$ paths and the total number of paths in the system is $O(n^2)$. The crossbar is more efficient in that each processor must support only 1 path. However the crossbar

must have n^2 switching points and a processor can be connected to only one other processor at a time.

Other classical interconnection arrangements which limit the number of interconnections (and thus the interconnectability) yet retain some performance are the nearest neighbor mesh and the ring. In the mesh, processors are arranged in a logical grid with each processor bi-directionally connected to its nearest neighbors. Processors along the edges of the grid may be connected to processors on the opposite edge forming a closed network. The mesh connected architecture is generally employed on systems with a large number of processors. The ring network has each processor connected with two neighbors and the entire system configured as a linear string whose ends close on each other. Data flow in a ring may be unidirectional or bidirectional. (There are some systems which are open rings.) In both the mesh and ring, a processor may exchange data directly with its neighbors but data for any other processor must pass through one or more intermediate processors. In the ring the number of intermediate processors could become large causing significant delays. In a practical sense, a balance between the number (thus cost) of paths provided vs the delays incurred by a less than optimum capability must be achieved.

All the architectures listed above are primarily message passing oriented though the crossbar scheme has been also used in a shared memory architecture. There the crossbar switch is used to dynamically connect different processors to multiple memory banks.

An important development in the evolution of shared memory systems came with the development of Cm* at Carnegie-Mellon University [3]. Two important advances in architecture were embodied in this system. First was the ability for processors to access the memory space of other processors without the necessity of switching mechanisms. This was accomplished by providing hardware which mapped part of a processor's address space into the physical memory of another processor. The second important development of Cm* was its hierarchical nature. The architecture consisted of a number of Cm's (computer modules) which were organized into clusters of 10 Cm's each. The Cm's communicated with each other over a local bus (cluster bus). Five clusters were then interconnected via a second intercluster bus to form a second level of the hierarchy. Within the system any Cm could access the memory of any other Cm. Thus data movement within one cluster does not interfere with data movement in another cluster.

19.3.3 Recent Multiprocessor Architectures

The current generation of microprocessor systems has opened the way for the cost-effective implementation of multiprocessor systems. From the available technology, a number of architectural options are available. To optimize the design and efficiency of such systems, a thorough understanding of the characteristics of variations in architecture is necessary.

Microprocessor design has facilitated the construction of multiprocessor systems. Along with the microprocessors, several standardized bus designs have evolved which allow the interconnection of system components. Foremost among these bus standards are the Multibus and the VME bus. When processor boards, memory boards, and input/output boards are designed to be compatible with one of these standard buses, a complete computer system can easily be configured. These bus standards allow multiple processors to reside on the bus and share resources (memory and I/O). Processor modules may have local memory. Given proper processor board design, the processors have the capability to read and write each others memory. These designs allow multiprocessor systems to be readily constructed. Figure 2 depicts the block diagram of a typical bus based multiprocessor system.

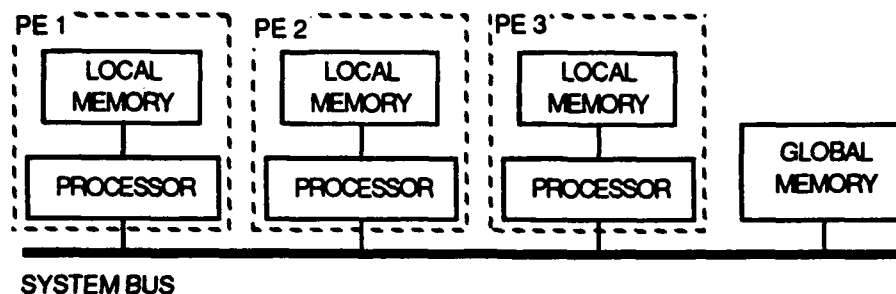


Figure 2: Bus Based Multiprocessor System

Another recent development is the hypercube architecture. This architecture attempts to minimize the number of paths required along with minimizing the number of intermediate processors a message between processors may have to pass through. For an n -processor system (n is generally a power of 2), this architecture requires each processor support $m = \log_2 n$ paths and a message will have to pass through at most $m-1$ intermediate processors. The hypercube and the ring architectures are both suited for real-time control applications since they work well with small numbers of processors. Figure 3 shows the interprocessor communication paths for an 8-node hypercube.

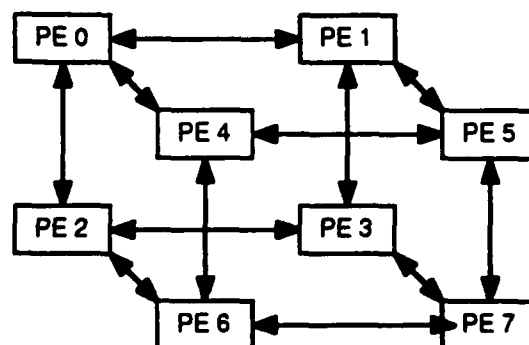


Figure 3: Hypercube Multiprocessor

19.3.4 Hierarchical Architectures

The Cm* was a hierarchical architecture in that there were two levels of interconnecting buses. However in this system all processors were peers. A logical extension of the Cm* hierarchical architecture was to place processing elements at more than one level of the hierarchy. Several research systems have been developed employing this concept. Most of these systems utilize a group of identical processors. Following in this direction, research at the PEC has led to the development of systems which not only have a multilevel hierarchy of processors but allow different types of processors to be intermixed.

This development has led to the Heterogeneous Hierarchical Architecture for Real-Time (H²ART). This architecture is characterized by a multi-level arrangement of processors and buses which are not necessarily identical. Like the architecture of Cm*, these systems can be structured such that there are clusters of interconnected computational elements (nodes). Then multiple nodes can be grouped together to form even larger computing platforms.

A H²ART system node consists of one to three computational processors and a node controller. These processors are interconnected by a local bus. A larger system can then be structured by interconnecting multiple nodes across another bus. Communication between computational processors occurs under control of the node controller. Likewise messages can be passed between computational processors on different nodes across the interconnecting bus under control of the node controllers. The entire H²ART system is supported by a host system which communicates with the nodes across the system bus. Figure 4 depicts the architecture of the H²ART system.

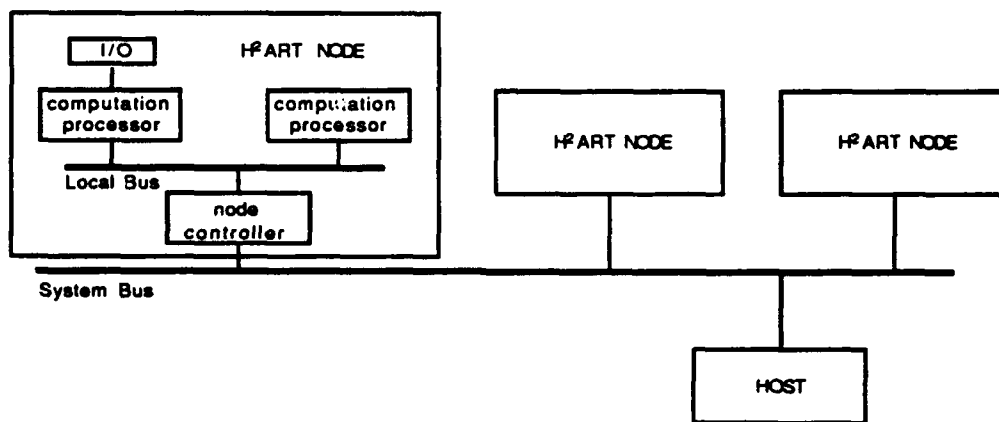


Figure 4: H²ART Multiprocessor System

H²ART architecture forms the basis of three multiprocessor systems currently under study at the PEC. Each of these systems employs an Intel 80186 processor as a node controller and multiple digital signal processors (DSP) as computational elements. The DSP's employed in these three

systems are the Texas Instruments TMS320C25, the TI TMS320C30, and the AT&T DSP32. Multibus-I provides residence and communications for the nodes. At present, a PC-AT serves as host for these systems. Additionally a commercially available H²ART architecture system is available for study. This system employs a Motorola 68030 as node controller and TMS320C30's as computational processors. This system resides on a VME bus and will be hosted by a SUN Workstation. While possessing a similar overall architecture, each of these systems utilizes a different interprocessor communication scheme thus providing a vehicle for studying the performance aspects of differing architectures. The understanding of the strengths and weaknesses of each of these is part of the research focus.

19.4 ARCHITECTURES OF PEC RESEARCH SYSTEMS

The current research activity on the hardware aspects of interprocessor communications is directed toward characterization of the communications performance of the available H²ART architectures. Additionally these same characterizations will be done for a conventional Multibus based multiprocessor using both memory mapping and separate global memory. These characterizations will provide the basis for development of performance models for these architectures. From these evaluations, it is expected that conclusions can be derived regarding the architectural direction new systems should take to provide for highly efficient interprocessor communications.

To study the performance impact of different architectural designs on communications, five multiprocessor systems are being studied. The following paragraphs briefly describe each architecture and the communication facility available with the design.

19.4.1 Multibus Based Multiprocessor

This system is a conventional homogeneous, bus based multiprocessor [9]. The evaluation of this architecture is important since it represents a baseline conventional multiprocessor system. With such a system, three different communications mechanisms are being studied. These are:

1. Use of shared global memory as a message buffer and global data area.
2. Allow any processor to directly access memory areas within any other processor
3. Use one processor as a communications hub through which all messages are passed.

The particular hardware being used for this architecture is a set of Intel 86/30 processor boards residing in a Multibus chassis. While the 8086 processor on these boards does not represent the latest in microprocessor technology, the system is quite adequate for evaluating the data movement characteristics of this type of system architecture.

19.4.2 TMS320C25 H²ART Multiprocessor

The C25 H²ART node consists of an Intel 80186 processor as the node controller and one to three Texas Instruments TMS320C25 Digital Signal Processors as computational processors [10]. Communications between computational processors can be accomplished using either point-to-point serial ports (one input & one output on each DSP) or across the local bus with the 80186 controlling the data movement. In this architecture, each computational processor has two independent 4K byte banks of data memory which are accessible by both the computational processor (C25) and the node controller. Access to these banks of memory is mutually exclusive. This allows the computational processor to be working out of one bank while the node controller is moving data to/from the other bank. This avoids delays due to memory access contention. Data movement across the local bus does not require participation by the computation processor. The C25 nodes interconnect via Multibus.

19.4.3 TMS320C30 H²ART Multiprocessor

The C30 H²ART node consists of an Intel 80186 processor as the node controller and one to three Texas Instruments TMS320C30 Digital Signal Processors as computational processors [11]. This architecture is essentially identical to the C25 H²ART node except for the common memory access by the C30 and the 80186. Here the computation memory is accessible by both the DSP and node controller via a true dual port memory interface which allows simultaneous access to any word without contention. This capability eliminates the C25 system's need for bank switching when data is to be moved between processors. The C30 nodes interconnect via Multibus.

19.4.4 DSP32 H²ART Multiprocessor

The DSP32 H²ART node also employs an Intel 80186 processor as the node controller and three AT&T DSP32 Digital Signal Processors as computational processors [4]. Communications between computational processors can be accomplished using either point-to-point serial ports (one input & one output on each) or across the local bus with the 80186 controlling the data movement. The DSP32 provides on-chip I/O registers which allow an external processor (80186) to command I/O transfers. Thus data movement across the local bus does not require participation by the computation processor. The DSP32 nodes interconnect via Multibus.

19.4.5 PC/M DSP-2A Multiprocessor System

The DSP-2A multiprocessor is a commercially developed system which follows the H²ART architecture [5]. It consists of a Motorola 68030 processor as the node controller and three Texas Instruments TMS320C30 digital signal processors as computational processors. These nodes reside on a VME bus. This system does not provide external point-to-point serial ports like the

other H²ART architectures being studied. This architecture provides three different interprocessor communication facilities. These are:

1. 512K bytes of global memory all of which is accessible by all four processors and part of which is accessible across the VME bus.
2. Bidirectional FIFO buffers between the three C30's.
3. A hardware mailbox scheme.

The DSP-2A's variety of interconnection capabilities provide the ability to choose the best communication mechanism for the application.

19.5 EXPERIMENTAL RESULTS

To date, extensive performance measurements have been performed on the DSP32 based multiprocessor system. These measurements have provided an accurate modeling of the communications performance of the DSP32 H²ART system. In addition to communication performance measurements, this system was used to implement a demonstration of a hypothetical system model used in the theoretical work to demonstrate deadline failures in a pipelined system.

19.5.1 Test Environment

Performance measurements and demonstrations were run on a two node DSP32 based multiprocessor system. Each DSP32 node physically consists of a single Multibus form printed wiring board. A PC/AT serves as the system host and interfaces to the multiprocessor system via a BiT3 PC to Multibus interface. The two nodes and the Multibus end of the Bit 3 adapter are housed in a Multibus chassis.

For measurement of the system timing the hardware is instrumented with oscilloscopes and a specially designed hardware measurement unit. These devices are attached to selected internal and external signals and are used to measure the duration and time relationship of system activity. The oscilloscopes are used to observe the relationships of the various marker events and to determine gross timing values. The hardware measurement unit provides the ability to count and time events with a 100 ns resolution. This device is used to obtain accurate measurements of timings. One limitation of the hardware measurement unit is that it obtains the average timing of a large number of repetitions of the same event but does not provide detailed event-by-event analysis such as minimum/maximum duration, histograms, etc. An advanced unit which can provide this level of detail is being proposed.

The hardware signals used as event markers are generated in two ways. First, some signals are created by the microprocessor or DSP's in the normal course their operation. Secondly, other

signals are explicitly generated under software control. Where software commands are used to create marker signals, care was taken to minimize the impact of such embedded instructions on the measured performance.

19.5.2 Communications Blocking Demonstration

To demonstrate the impact of task blocking due to synchronous communications, the hypothetical four processor pipeline example used in the description of the theoretical development was implemented. This example illustrated that variation in the computation time of a task other than the critical task of the pipeline could cause a deadline failure of the critical task. The critical task is the task with the longest cycle time. Figure 5 shows the decomposition of the hypothetical example.

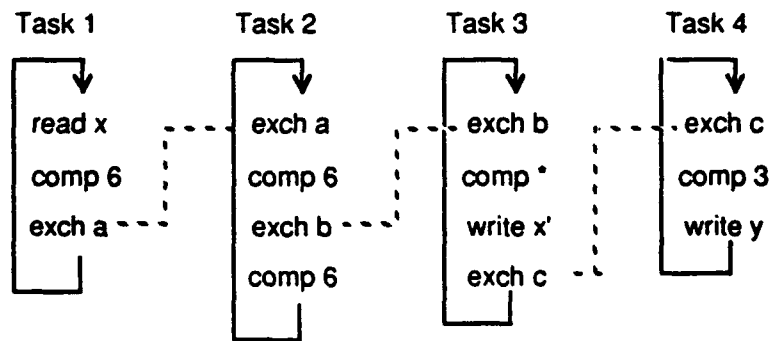


Figure 5: Four Task Decomposition of Pipelined Application

The data exchange was assumed to require 2 units of time and the read/write assumed to take 1 unit. Thus the individual task cycle times for tasks 1, 2, and 4 are 9, 16, and 6 units respectively. Task 3's computation time (C_3) varies from 1 to 10 time units giving it a cycle time of 6 to 15 units. The pipeline cycle time is then set by the critical task's cycle time (Task 2 = 16 units). Tasks were assigned to computational processors (DSP32's) as shown in Table 1.

Task	Node	DSP
1	0	1
2	1	1
3	0	2
4	1	2

Table 1: Task Assignment for Communications Blocking Demonstration

Interprocessor message length was chosen to be 16 bytes giving a message transmission time of ~3.66 ms. Therefore 1 time unit of the hypothetical example equals ~1.83 ms. DSP programs were created to emulate the computation times and communications of the example. A series of

tests were run varying the computation time of task 3 (C_3) from 1.8 ms to 18.3 ms (1 to 10 time units). In these tests, the pipeline cycle time remained relatively constant (~31.9 ms) for values of C_3 up to 3 time units (~6 ms). Then the cycle time begins to rise as task blocking occurs. This occurs at the equivalent of 4 and 5 time units. Then at the equivalent of 6 time units the cycle time drops back to the non-blocking level. In this demonstration, the worst case pipeline cycle time outside the region of blocking is 31.9 ms which implies that the execution time of the critical task (including communications) is 31.9 ms. The equivalent delay times and the measured pipeline cycle times are summarized in Table 2. Figure 6 is a graph of these results.

C_3 (units)	C_3 (ms)	Cycle (ms)
1	1.8	32.3
2	3.7	31.3
3	5.5	31.9
4	7.3	33.9
5	9.2	35.6
6	11.0	30.1
7	12.8	31.3
8	14.6	31.2
9	16.5	31.2
10	18.3	30.9

Table 2: Blocking Delay Demonstration

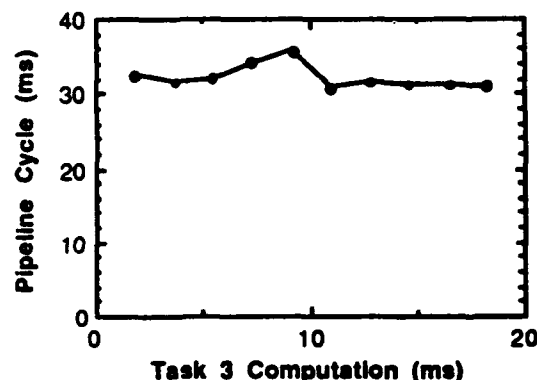


Figure 6: Blocking Delay Demonstration

The theory says to add the worst case communication time for each use of the communications resource to the longest task's execution time to get the minimum cycle time which can be guaranteed. In this example this is 31.9 ms plus 7.2 ms (3.6 ms x 2) which gives 39.1 ms as the minimum cycle time which can be guaranteed. The observed worst case cycle with blocking occurring (35.6 ms) is less than this bound. This demonstrates that application of the theoretical methodology for setting the minimum cycle time which can be guaranteed does work.

19.5.3 DSP32 Communications Performance

To evaluate the performance of communications on the DSP32 multiprocessor system, tests were run on the two node system described above. Test scenarios consisted of passing messages of various lengths between pairs of DSP's. Message lengths were varied from 4 bytes up to 2000 bytes. Transfer times between DSP's on the same node and DSP's on different nodes were measured. The results of this investigation showed a rather inefficient performance. The interprocessor data transfer rates for small messages (<100 bytes) are shown in Table 3.

	Setup time (μ s)	Transfer time (μ s/byte)
Intranode	3090	55
Internode	2257	84

Table 3: DSP32 Multiprocessor Communications Performance

Thus a message consisting of 10 floating point data words (32 bit) takes approximately 5.3 milliseconds to pass from one DSP to another DSP within the same node and 5.6 milliseconds when passed across the Multibus to a DSP on a different node. This speed is unacceptably slow for high-performance real-time systems.

With such poor throughput, the communications mechanism was analyzed in more detail to determine why it was so slow. This analysis showed that the data transfer interface between the DSP32 computation processor and the 80186 node controller is the bottleneck. Even though the DSP32 has an on-chip DMA capability and the node controller also has DMA capability, these were not fully exploited in the design of the system. As a result there is a considerable amount of handshaking required for the node controller to transfer a data word to/from the DSP. Even if the DMA capability was utilized, it would not greatly improve the performance of the many short (1-4 16-bit word) housekeeping messages which are exchanged as part of the required communications protocol. For example, a 2-word message from DSP to 80186 requires 206 μ s and a 2-word message to the DSP from the 80186 requires 224 μ s. Since there is no interrupt structure implemented such that the DSP can signal the node controller when it has a message to send (or wishes to receive one), the node controller must continuously poll the DSP's to see if any message send/receive requests are pending. The result of this is that the polling loop alone requires a minimum of 618 μ s (actually measured at 656 μ s).

As a result of the in-depth study of DSP32 H²ART's communications, some preliminary results relative to the performance of a more conventional multiprocessor system employing a common bus were obtained. Specifically measurements of the time required to pass messages across the Multibus were obtained. These results showed a data transfer time of 7 microseconds/byte. This implies that a message of 10 32-bit words could be passed between processors in ~300 μ s.

The serial links of the DSP32 H²ART operate at a data rate of 2 MHz. This means that transfer of a 32-bit floating point data value can occur in 16 μ s. Though the serial links have the potential of faster transfer than the parallel path through the node controller, they are limited. Since serial links are point to point, they must be preconnected between specific DSP's. This eliminates the ability of a DSP to send or receive messages to or from more than one DSP. This could pose a serious limitation in real-time control applications.

19.6 SUMMARY AND CONCLUSIONS

This section has reported the results of continuing work at the PEC related to the development of methodologies for design of real-time control systems with guaranteed performance characteristics. The results of an analytical study of the impact of interprocessor communications over a common communications resource were summarized. The theoretical work initially focused on pipelined applications and has now been extended to include the general case of shared resources in a multiprocessor system. The evolution of multiprocessor architectures with a primary focus on interprocessor communications was then reviewed. This background establishes the origins of the multiprocessor architectures developed. Five multiprocessor systems that are available for experimental evaluation were described and the results of the experimental studies conducted were reported. These experimental studies included a demonstration of the application of the theoretical results to a specific case and an in-depth evaluation of the communications performance of one of the five research multiprocessors.

While measurement of communications performance of the available multiprocessor architectures is incomplete, the data already available indicates some important conclusions. Completion of the experimental phase of this project is expected to confirm these preliminary conclusions.

One conclusion is that having the node controller poll the computational processors to determine if a communication request is pending is undesirable. Polling for message send/receive requests has two negative impacts of system performance. First the polling cycle time represents an additional overhead on the message passage time. Secondly, to maintain maximum responsiveness the node controller cannot do any other tasks except communications serving.

Additionally, performance is very dependent on the amount of handshaking which must occur between the processors to pass a message. This is both a hardware and software issue. A fast data transfer capability may not be of great advantage if the many handshaking/setup messages must be passed before the actual data message is sent. In most multiprocessor control applications, the messages exchanged between processors are generally quite short. Thus the hardware/software architecture should be optimized for minimum overhead.

19.7 FUTURE WORK

This project continues to investigate the communications performance of various multiprocessor architectures. Short term goals are to conduct measurements similar to those done on the DSP32 H²ART system on the C25 and C30 H²ART systems and the 86/30 Multibus system. From the results of these investigations a recommendation will be developed for an interprocessor communications architecture for H²ART systems which will be both efficient and cost-effective.

References

- [1] Downey, J.K., "A Petri Net Based Timing Analysis Tool for a Real-Time Multiprocessor Integrated Programming Environment", M.S. Thesis, Department of Computer Science, North Carolina State University, Raleigh, NC, 1991.
- [2] Fornaro, R.J. & Allen, W. D., "Application of Real-Time Scheduling Theory to Multiprocessor Pipelines", *Proceedings: Eighth IEEE Workshop on Real-Time Operating Systems and Software*, May, 1991.
- [3] Gehringer, E. F., Siewiorek, D. P. and Segall, Z., *Parallel Processing: The Cm* Experience*, Digital Press, 1987.
- [4] National Security Agency, *DSP32 - Multibus Computer User's Manual*, June 1989.
- [5] Pacific Cyber/Metrix, *DSP-2A: Multi-DSP Processor User Manual*, September 1990.
- [6] Sha, L., Rajkumar, R. and Lehoczky, J. P., "Priority Inheritance Protocols: An Approach to Real-Time Synchronization", *IEEE Transactions on Computers*, Vol.39, No.9, September 1990.
- [7] Skroch, D.A., "A Hierarchical Architecture for Real-Time", M.S. Thesis, Department of Electrical & Computer Engineering, North Carolina State University, Raleigh, NC, 1989.
- [8] Smith, M.B., "SIMPLE: A Multiprocessor Programming Environment For Real-Time Applications", M.S. Thesis, Department of Computer Science, North Carolina State University, Raleigh, NC, 1990.
- [9] Taylor, L. W., Fornaro, R. J. and Garrard, K. P., "The Architecture of an 8086 Based Multiprocessor", Technical Report TR-86-22, Department of Computer Science, North Carolina State University, Raleigh, NC, 1986
- [10] Taylor, L. W. and Fornaro, R. J., "Development of Prototype Multiprocessor Architecture", *Precision Engineering Annual Report*, North Carolina State University, Raleigh, NC, Vol. VII, pp. 1-7, December 1989.
- [11] Taylor, L. W. and Fornaro, R. J., "Development of Prototype Multiprocessor Architecture", *Precision Engineering Annual Report*, North Carolina State University, Raleigh, NC, Vol. VIII, pp. 319-324, December 1990.

20 PERFORMANCE MEASUREMENT OF UNIX AS A REAL-TIME OPERATING SYSTEM

Dwayne Allen Sorrell

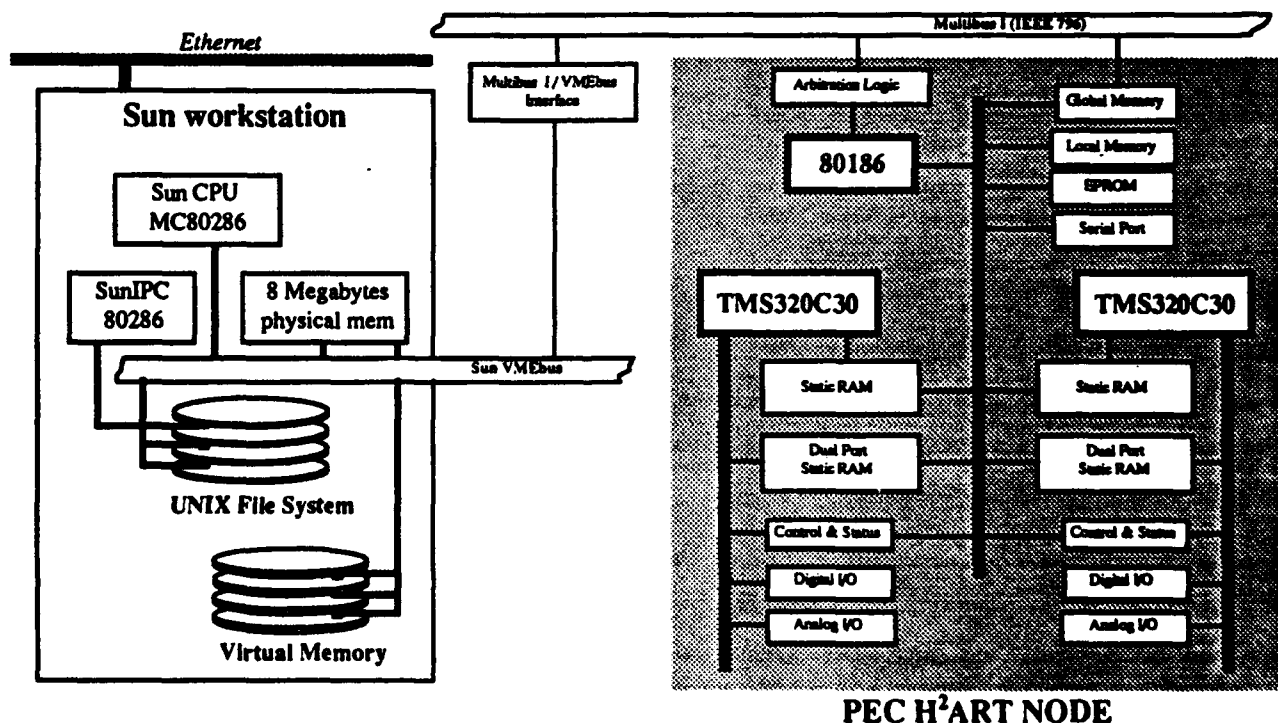
Graduate Student

Robert J. Fornaro

Professor

Department of Computer Science

The UNIXTM operating system provides for a powerful user interface and development system. However, UNIX is a time-sharing system and was not designed for real-time applications. By integrating the H²ART as a real-time embedded system, UNIX can be adapted for use with real-time applications, but only in conjunction with an embedded real-time system. In this hybrid mix, the performance of the entire system is controlled not only by the speed of individual processors within H²ART and UNIX systems, but also by the communication performance between systems. Benchmark utilities have been developed to better understand and characterize communication between H²ART processor memory and UNIX. Accurate communications measurements give the real-time application developer a guide to expected system performance as well as system limitations.



20.1 INTRODUCTION

Current implementations of the H²ART system are hosted by a PC/AT compatible computer. One of the goals in this research project is to use a UNIX workstation as the user interface and development system for H²ART systems. This would combine all the benefits of a UNIX workstation (graphics, multi-user, networking, software engineering tools, large file system, etc.) with the real-time power of the H²ART. The result will be a powerful UNIX operating system extended to include real-time capabilities. There are commercially available real-time systems which provide for kernels on an embedded real-time system as well as development tools for the UNIX host. However, there are no known alternative systems that can provide the level of functionality required to meet the real-time demands of precision engineering. In general, commercial real-time systems base their performance capabilities on a guaranteed response time for asynchronous interrupts. This means that the system will be able to respond to non-periodic hardware events within a predicted (pre-measured) amount of time. This does not mean that the system is capable of completing a task within a required deadline. Many control applications are periodic with the deadlines for task completion related to input frequency. Another common characteristic of commercial systems is indirect I/O. In these systems, the data movement is controlled at a high level by non-real-time components and then routed to and from dedicated real-time processors. This I/O latency was avoided in the H²ART system by allowing the DSPs at the lowest level direct access to I/O.

This research effort to combine strengths of a UNIX workstation with H²ART will soon be able to provide the real-time functionality needed for precision engineering applications. This cannot be achieved under UNIX directly. Performance levels required for precision engineering applications may be provided indirectly by measuring UNIX performance capabilities, relating performance to values of system parameters, and guaranteeing response and throughput by selecting appropriate values of system parameters. These measurements provide the application developer a guide to expected real-time performance as well as system limitations.

Communication has been established between a Sun 3/260 and a H²ART system. By using a Bit 3 bus adaptor, the H²ART's Multibus is connected to the Sun's VMEbus and communication is established by using physical bus addresses on both sides. This technique has been in use for some time with PC to H²ART communication. To date, the primary effort was on getting the Sun workstation and H²ART system to communicate. Now the focus has shifted to measuring the communication performance between systems and developing guidelines for real-time application development. Another important goal of this project is to facilitate the transition of the H²ART/C30 processor from the 16 bit Multibus environment to the 32 bit VMEbus environment. It is expected that the Sun workstation will be able to perform as a user host and development system and perhaps replacing the existing PC-H²ART configuration.

20.2 UNIX-H²ART DEVELOPMENT

Development tools for the UNIX-H²ART system have been written as the need arises. The tools fall into three categories: software development tools such as compilers, interface tools to communicate with the H²ART, and performance tools for testing the capabilities of the H²ART system and the communication between host and H²ART.

20.2.1 Software Development

Currently, most software development tools for the Intel 80186 node controller and the TMS320C25 and C30 DSPs reside only on the PC. While the compilers do exist for the SunOS environment, their cost is prohibitive. Instead, there are two alternatives which can be used. Code can be compiled for the processors on a remote PC and later moved to the Sun before downloading. Also, the Sun contains an internal 80286 PC (SunIPC) that can be used to create and compile code directly from the Sun console. This second alternative has proven to be much more convenient. Both UNIX and DOS are accessible to the programmer from the Sun console, and the internal PC shares the UNIX file system with the Sun. The internal PC is accessed through a window on the console. This configuration allows the developer to simultaneously use all the Sun's UNIX utilities as well as all the PC's compilers and DOS utilities.

20.2.2 Interface Tools

The main interface tool used on the PC hosted H²ART systems is MPC [1]. This program has been replaced by *hh* [2]. Initial work in porting *hh* to the Sun is in progress. In the interim, existing tools, such as MPC, as well as *iod* and *download* [3], will be used as the primary interface tools.

- *iod* is designed as an interactive patch utility. It can also be used to produce dumps of Sun or H²ART memory and files. It is similar to the UNIX *od* command but with a more flexible input and output formats for easier use.
- *download* provides for the downloading capability needed to load UNIX disk files and 80x86 executables into H²ART memory.

20.3 PERFORMANCE ANALYSIS

UNIX is a time-sharing system and therefore not designed for real-time applications. By integrating the H²ART as a real-time embedded system, UNIX can be adapted for use as a real-time operating system. The H²ART system by itself is a real-time system with well defined performance and timing characteristics. By combining H²ART with UNIX, the UNIX system

acquires the real-time capabilities of the H²ART , but the H²ART system suffers in its real-time performance whenever it must communicate with the less predictable UNIX host. The communication performance can be characterized and measured with the use of well constructed benchmark facilities.

Recent work has been concerned with gaining a better understanding of how to transfer data between systems as efficiently as possible. One simple approach is to copy data between systems and measure the amount of time it took. Initially, a program called *memtest* was written to do this testing. However, UNIX systems are quite complex in comparison to PCs. The time required to move data between the host system and the embedded system can depend on many different factors:

H²ART characteristics affecting communication bandwidth:

- 1 . Data transfer speed of bus interface.
- 2 . Memory speed.
- 3 . Synchronization and arbitration
 - Among processors with H²ART system.
 - Between host and H²ART

Host system characteristics affecting communication bandwidth:

- 1 . CPU activity
 - System load for the entire system.
 - CPU utilization of the real-time process.
- 2 . Buffer length
 - Size of data transfers.
 - Size of disk cache.
- 3 . Efficient use of virtual memory.
- 4 . Disk activity
 - System wide disk usage.
 - Duration and type of disk usage by real-time process.
 - Synchronous disk writes.
- 5 . Using contiguous files versus a UNIX file system organization.
- 6 . File system parameters.
- 7 . Disk geometry.
- 8 . Disk controller.

Memtest was able to assist in verifying the system characteristics that most affected real-time performance. However *memtest* was not very sophisticated and could not accurately measure the

performance effects of each of these system characteristics. As a result, a new benchmark program called *commbench* was developed to better study these effects.

20.3.1 Communication Benchmarks

The general definition of a computer benchmark is a standard program used to measure the relative performance of one computing system in comparison to another. Examples of this type of program would include WHETSTONE [4], DHRYSTONE [5], and the SPEC benchmark suite [6]. However, benchmarks can also be used to measure the absolute performance of a computing system. It is this type of absolute performance analysis that is being described in this report. Most of the available benchmark programs, commercial and public domain, are comparative benchmarks and therefore are not of much use to this research effort. A few benchmarks such as *iozone* [7] and *Rhealstone* [8] are designed for absolute performance analysis, but are far too restrictive in their capabilities to be of any use. *Iozone* can only handle certain types of file operations and only measures data transfer speeds. The *Rhealstone* benchmark was designed to measure the following six operations.

- 1 **Task-switch time:** average time to switch between two tasks of equal priority.
- 2 **Preemption time:** average time for a higher priority task to preempt a lower priority task.
- 3 **Interrupt latency:** average delay between an interrupt event and the execution of the first application-specific instruction within the interrupt handler.
- 4 **Semaphore shuffle time:** delay between the release and acquisition of a semaphore when at least one task is waiting on the semaphore.
- 5 **Deadlock-break time:** average time to break a deadlock caused by a lower priority task holding a resource required by a higher priority task.
- 6 **Intertask message latency:** average time to send a message from one task to another.

The authors considered these operations "vital indicators of real-time multitasking system performance" [8]. *Rhealstone* is not designed to measure general system communication. With no known alternatives, a customized benchmark utility called *commbench* was created to characterize H²ART and Sun performance. *Commbench* is a generic communications benchmark utility for measuring system resources required to move data from point A to point B. This includes data movement within the host as well as between the host and embedded system. The advantages of *commbench* over its predecessor *memtest* are greater control over data transfer, very accurate measurements of system resources, a simple front-end to allow for more complicated benchmarks and more sophisticated statistical measurements. Figure 1 shows the proper syntax and which parameters can be controlled. Default values are shown in brackets.

Usage: `commbench [global_options] [-s source_description] [-d dest_description]`

Global options:

-h -help Print this help information and quit.
 -l n Data length to be transferred on each iteration. [0x80000 Bytes]
 -b n Data transfer block size. [file->st_blksize]
 -i n Number of times to repeat the data transfer. [100]
 -p n Boundary for memory allocations. [8192 Bytes]
 -n n Nice value (priority) to run benchmark. [0]
 -t n Test method used on memory transfers. [2]
 -O [n] Remove loop overhead from elapsed time. [0] 0=no 1=yes
 -F form Output status format string. [%\$BB. %\$BE.\n]
 -f file Output status format file name.

Source and destination description: `target[,option][,option...]`

target is 'physical' 'memory' or 'file'. [memory]

options: (separated by commas)

name=filename Target file name. [commbench.in & commbench.out]
 addr=starting addr Offset from beginning of file or memory buffer. [0]
 size=memory size Size of target address space. [Inf]
 access='sequential' | 'random' | 'mapped'. [sequential]
 sync Use synchronous write operations. [no sync]

Figure 1: *Commbench* syntax.

- Data flow rate.
- Elapsed (wall clock) time.
- Kernel CPU time spent for user process.
- User CPU time.
- Percentage of total CPU time used; user and system CPU time.
- Integral resident set size (unshared memory).
- Maximum real memory used; maximum set size.
- Average amount of shared memory used.
- Average unshared data size.
- Maximum unshared stack size.
- Average total memory usage.
- Number of block input and output operation to file system.
- Minor page faults; reclaims, no physical I/O.
- Major page faults; required physical I/O.
- Number of times swapped out.
- Signals delivered.
- Socket messages sent and received.
- Voluntary and involuntary context switches.

Table 1: Output values from *commbench*.

Unlike many other benchmark utilities, such as *iozone* described above, *commbench* was designed to do much more than just measure communication bandwidth. Simple bandwidth graphs are not enough to understand which system characteristics were influencing the communication speeds. More system information needed to be collected. Therefore *commbench* was extended to measure system resource usage during the running of the benchmark. Table 1 shows an abbreviated list of the system resources that are measured during the benchmark process. By running simple controlled benchmarks that focus on a particular component of system, and then examining the system resource values against the data bandwidth, behavior of individual components of the system can be carefully characterized. After the individual pieces are understood, then this information can be brought together to characterize the behavior of the system as a whole.

20.3.2 Communication Benchmark Illustration

An essential operation performed in a control system is gathering of data. This may be for recording a series of measurements, tracking the history of controller decisions for later study, or the extreme case of taking a snapshot of a real-time processor and its local memory. Whatever the operation, there is a communication bandwidth that must be maintained between systems to guarantee real-time performance. In each of the example cases, if the data is large and cannot fit in the memory of the H²ART system, then the data must be transferred to the larger host system. If the data is extremely large and cannot fit in the memory of the host system, then the data must also be transferred to the host file system. What follows is a series of graphs which show an analysis of this process. This analysis has been simplified to illustrate system characteristics affecting communication performance and not necessarily maximum possible bandwidth. These benchmarks were run using default parameters shown in Figure 1. To achieve the maximum bandwidth, many system parameters would need to be tuned based on the results of several benchmark experiments.

Moving Data from H²ART to Host Physical Memory The simplest way to move data from H²ART system to the host system is to allow the host to copy an array from the H²ART's global memory in Multibus address space to an array in the host system. For all of these experiments, the host system is a Sun 3/260 with 8 Megabytes of physical memory. Figure 2 shows the results of simple benchmarks copying H²ART data to the Sun host. Each point represents a single benchmark run with the data length shown on the X-axis and the resulting data bandwidth shown on the Y-axis. For small amount of data, there is a substantial penalty due to a required startup time. Once the data size is at least 1 Kilobyte, this penalty is no longer a factor. At this point some other system characteristic becomes the limiting factor.

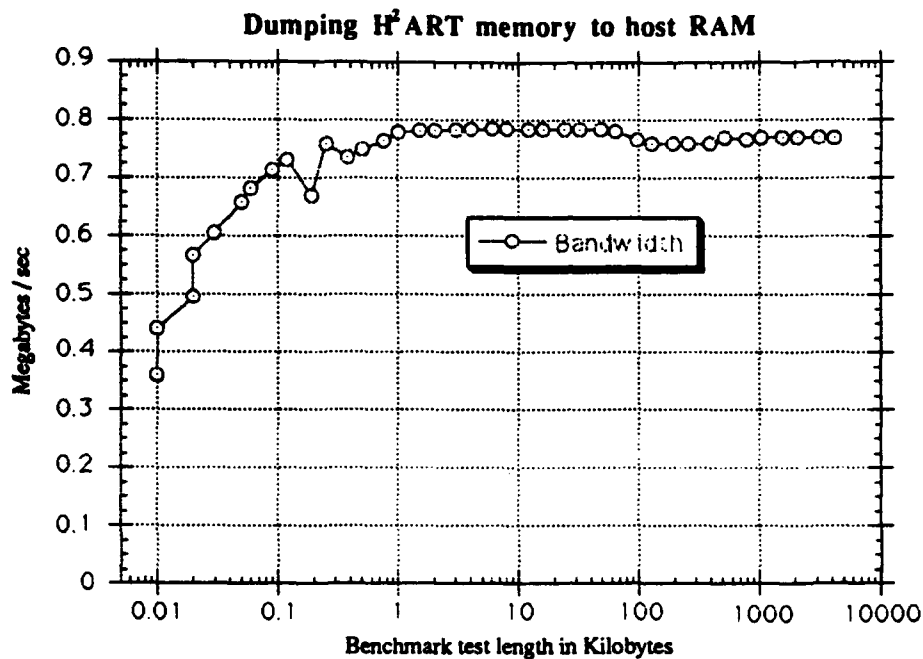


Figure 2: Copying H²ART memory to the host system memory.

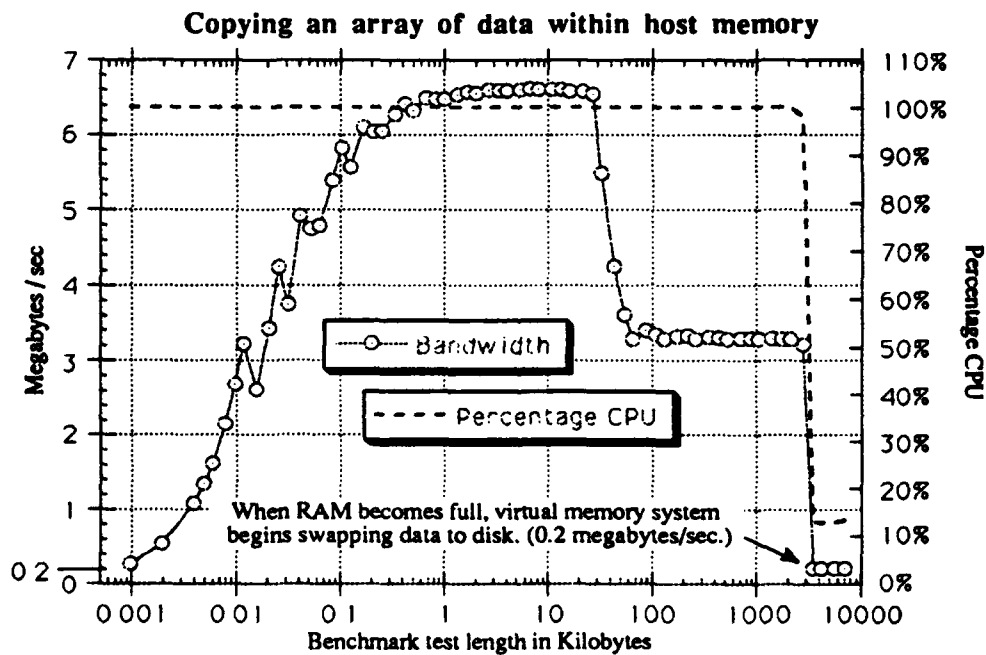


Figure 3: Copying memory within host system.

To determine which system (host, H²ART, or even the bus interface) is limiting the bandwidth to just less than 0.8 megabytes/sec, internal data transfers speeds can be examined separately in each part of the whole system. Figure 3 shows the results of moving data within just the host. As illustrated here, the host system can easily maintain an internal data bandwidth in excess of 3 megabytes/sec. Therefore, by comparing Figures 2 and 3, it can be concluded that the host is not responsible for the limited bandwidth. Either the bus interface, communication speeds within the H²ART system, or the combination of the two is limited to about 0.8 megabytes/sec.

Moving Data from H²ART to Host File System If the data is too large to fit in the host system's physical memory, then it must either be swapped to disk by the virtual memory system or it must be written directly to disk on the host system. As shown in Figure 3, if the data is allowed to be swapped to disk, then the performance should not be any better than about 0.2 Megabytes/sec. To understand what happens when writing H²ART data to a disk, it is important to first understand what happens when an array of data local to the host is written to disk. Figure 4 shows the results of a series of benchmarks in which an array of data is written to a new file in the host file system. For a small amount of data, there is a familiar penalty due to the required startup time. However in this case, performance suffers again after reaching the 8 K bytes block size. This is because up until the 8 K bytes limit, the system had not yet started writing anything to disk. This is evident by the dashed line representing the number of block operations for writing the data to disk. From a data length of 8 K bytes until 1 Megabyte, performance is at its lowest because the UNIX file structure is still being allocated. Afterwards performance levels off. Therefore, it is expected that if data is to be written from H²ART memory to the host file system, performance should be no better than the host system writing data to its own local file system. In fact, it is expected that the performance should reflect both the behavior of copying memory between systems, as in Figure 2, and writing a new file, as is Figure 4. This conclusion was demonstrated by the benchmark results shown in Figure 5. The graph of the bandwidth has the same general shape as in Figure 4, and the entire system is still limited to an upper bandwidth of about 0.8 Megabytes, which was measure in Figure 2.

20.3.3 Conclusions from Benchmarks

Figure 5 illustrates only a small part of the dataflow analysis. However within this one example several of the characteristics affecting bandwidth (listed in Section.3) were shown to be important factors. This example focused mainly on data transfer length as a significant variable affecting communication bandwidth. The expected performance for small data lengths will not be any better than about 0.8 Megabytes/sec, yet the host system is able to handle greater transfer speeds if the communication from H²ART to host can improve. For very large data transfers, the host system becomes a limiting factor. This measurement could also be improved by improving the host facilities. There are also many other variables listed in Section.3 that can be adjusted to improve performance.

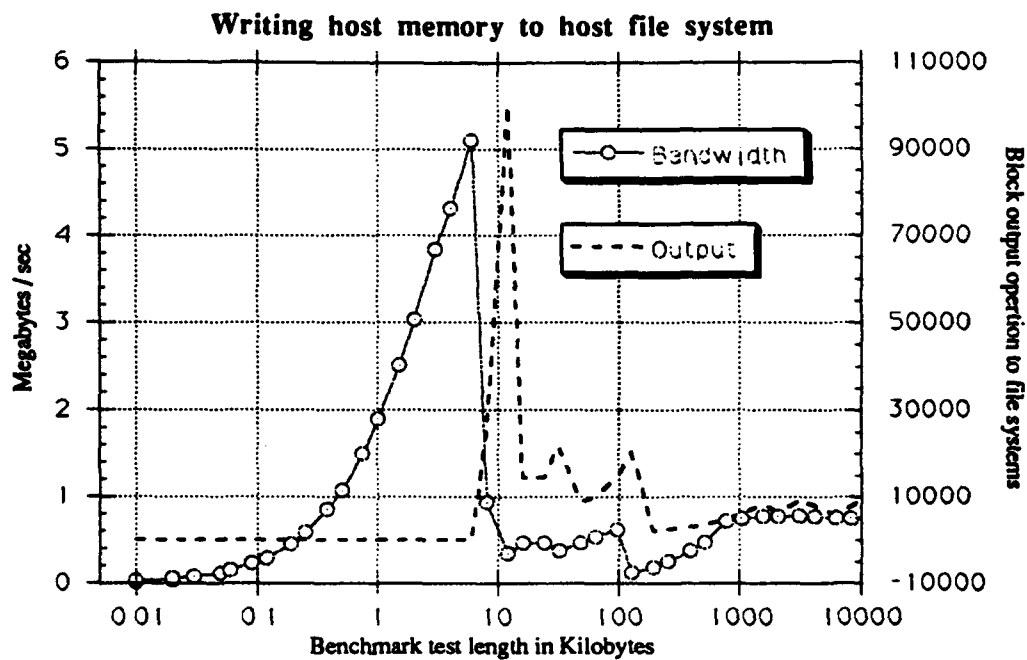


Figure 4: Copying host system memory to the host file system.

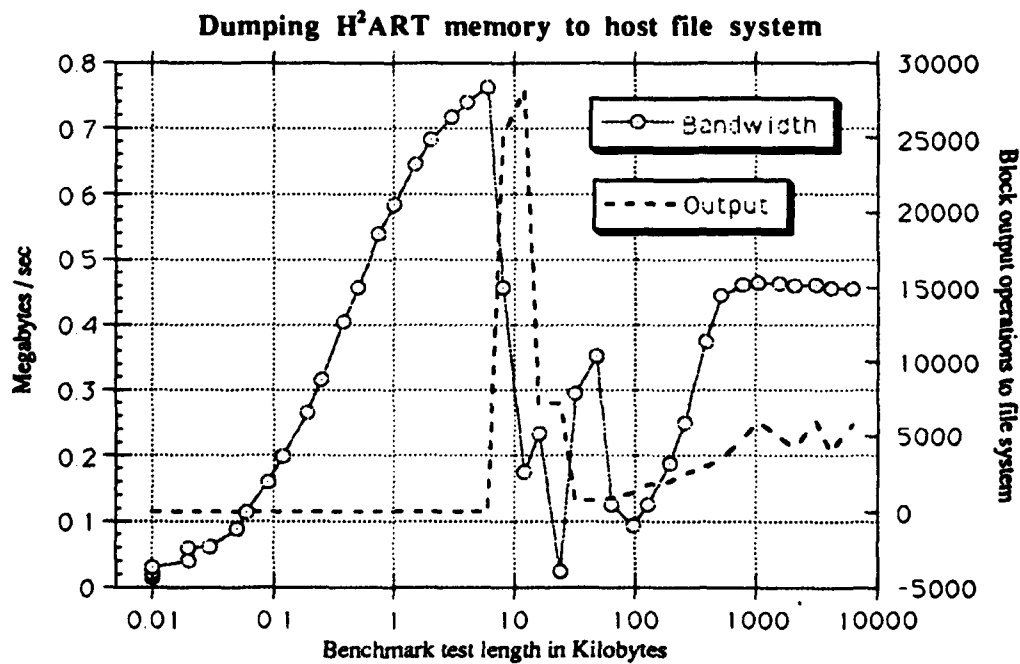


Figure 5: Copying H²ART memory to the host file system.

20.3.4 Performance Categories

Moving data from the H²ART system to the host is just one form of communication. Data may also need to be transferred in the opposite direction, from host to H²ART. For example, if the Fast Tool Servo is to cut a non-symmetric surface that requires a large number of data points, then the data will have to be stored in the host system and transferred to the H²ART node controlling the FTS during the cutting process. The performance characteristics of this type operation are quite different from the above example of dumping H²ART memory to the host file system. Figure 6 shows the expected performance when writing data from the host file system to H²ART memory. Instead of decrease in data bandwidth after the 8 K byte block size (as with writing to the file system), now the data bandwidth stays high when reading from the file system. This is because the UNIX file system attempts to improve performance by caching as much of the data as possible. The dashed line shows that system did not read data from disk until the test data was larger than the cache size. So there is an important difference in the performance of reading and writing operations.

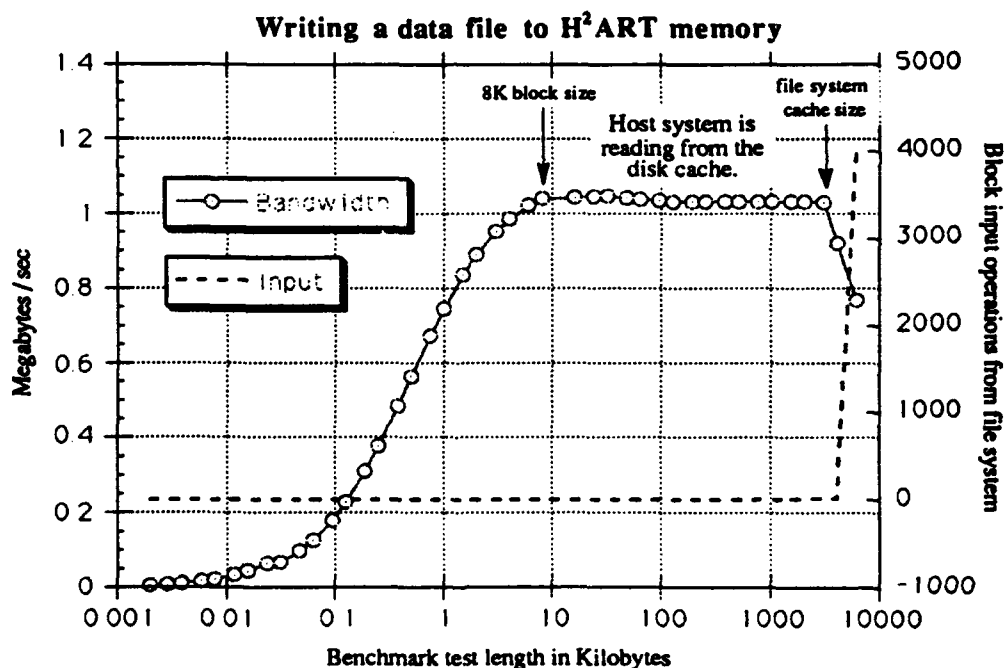


Figure 6: Writing data to H²ART from host file system.

By adding DSPs to the communication hierarchy, the analysis becomes more complicated. Nevertheless, the analysis methodology remains the same. As shown in Figure 6, when writing H²ART data to a UNIX file system, one-step communication is analyzed first. Later, communication steps are combined drawing conclusion from previous analysis.

20.4 FUTURE WORK

Commbench has been significantly improved since it was created in the Spring of 1991. There are still elements of system behavior that are not adequately explained by existing *commbench* analysis. In the past, whenever there was insufficient data for predicting communication performance, or the resulting performance could not be explained, improvements would be made to *commbench*. Therefore, there will undoubtedly be more enhancements to *commbench* as well as continued study of UNIX to H²ART communication. The next change will be to use a frontend to the *commbench* routines that will run numerous benchmarks in order to show the effects of modifying just one variable, such as data length. Figures 2 through 6 were created by running a large number of benchmarks. When the number of benchmarks in experiment becomes very large, it can take several hours just to create the data for one graph. Check points will also be added to the *commbench* frontend program. This will allow for check point monitoring of the experiment in addition to allowing the experiment to be temporarily suspended so that the Sun workstation can be used for other work.

The present mode of communication with the H²ART memory and nodes requires access through the UNIX file system. This creates a costly time latency that must be corrected for accurate communication timing analysis. Memory mapping to the Sun VMEbus is temporarily useful, however, a specialized device driver should be installed as soon as possible.

The host-H²ART interface tool *hh* will be ported to the Sun workstation. Once this is complete, the Sun will have the same H²ART interface capabilities as the PC system.

Work will also continue in the direction of graphics as a performance evaluator. This will provide for experimental verification of Sun-H²ART communication, *hh*, and benchmark results. Graphics programs lend themselves well to multitasking, and these routines can be collected into a library for future use in application of the H²ART systems.

References

- [1] Skroch, Denise A., "A Hierarchical Architecture for Real-Time," *M.S. Thesis, North Carolina State University, Raleigh, NC, 1989.*
- [2] Garrard, Kenneth P., "Diamond Turning Machine Controller Software Development," *Precision Engineering 1991 Interim Report, North Carolina State University, Raleigh, NC, 1991.*
- [3] Sorrell, Dwayne A., "Interprocessor Communication in Real-Time Operating Systems," *Precision Engineering 1990 Annual Report, North Carolina State University, Raleigh, NC, 1991.*
- [4] Wichman, B. A., "Validation Code for the Whetstone Benchmark," *Tech. Report NPL-DITC 107/88, National Physical Laboratory, Teddington, UL, Mar. 1988*
- [5] Weicker, Reinhold P., "Dhrystone: A Synthetic Systems Programming Benchmark," *Communication ACM, Vol. 27, No. 10, Oct. 1984, pp. 1013-1030.*
- [6] "Benchmark Results," *SPEC Newsletter, Vol. 1, No. 1, Fall 1989, pp.1-15.*
- [7] Norcott, Bill, Nov. 1989.
- [8] Kar, Rabindra P., "Implementing the Rhealstone Real-Time Benchmark," *Dr. Dobb's Journal, No. 163, April 1990, pp.46-55.*

21 AN APPROACH TO THE DESIGN AND ANALYSIS OF REAL-TIME COMPUTER CONTROL SOFTWARE

Andre N. Fredette

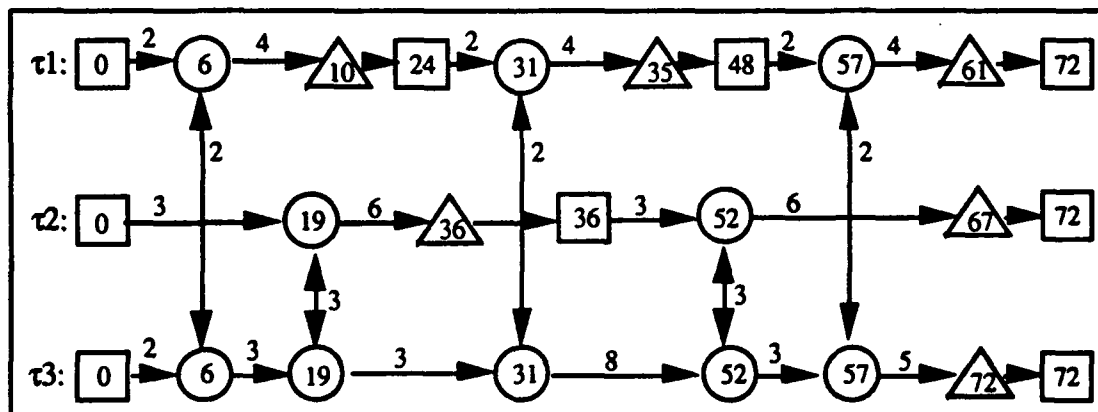
Graduate Student

Robert J. Fornaro

Professor

Department of Computer Science

Priority inheritance protocols have been shown to solve uncontrolled priority inversion problems which can arise in some applications of common synchronization primitives such as semaphores, monitors or the Ada rendezvous. However, existing results do not include systems in which tasks synchronize directly. It is demonstrated in this section that the basic priority inheritance protocol can be extended to allow schedulability analysis of tasks which communicate directly via synchronous message passing. The effects of synchronous message passing upon periodic tasks are outlined, an algorithm for accomplishing schedulability analysis of synchronizing periodic tasks is provided, and a method for the synthesis of minimal task periods is given.



21.1 INTRODUCTION

The scheduling of periodic tasks with hard deadlines equal to the task periods is a common problem in real-time computer systems. The periodic nature of these systems is due to the fact that the typical system must control a physical process. Most control algorithms require that various sensors of physical conditions be monitored periodically. The control system must synthesize the data it obtains during the monitoring process and respond accordingly. If a periodic deadline is missed, the consequences can range from a degradation of system performance to complete failure. In the case of precision machine tool control, for example, the results of missed deadlines can range from a figure error in a parabolic mirror to physical damage of the workpiece or machine tool. Since either outcome represents a total failure of the process, it is said that this application has *hard real-time* constraints. It is customary to conduct extensive pre-runtime schedulability analysis to guarantee that the system will not miss any deadlines. A set of tasks is said to be *schedulable* if it meets all specified deadlines.

It is possible to describe a control system in terms of functional modules. The modular description of the system can lead directly to a concurrent program, however, the use of concurrent programs can lead to several timing problems which will be analyzed in this paper. As described above, it is necessary to be able to make timing guarantees about the system as a whole. Each of the program modules can be individually analyzed for worst case timing performance. Conducting schedulability analysis of a sequential program is a non-trivial endeavor. When concurrency is involved (whether it be physical concurrency in a multiprocessor system, or logical concurrency in a multitasking uniprocessor system), the analysis becomes even more difficult. For example, the effects on system schedulability of different scheduling disciplines and the problems caused by the use of shared resources and inter-task communication must be considered. Methods for analyzing hard real-time computer systems which use fixed priority preemptive scheduling on a uniprocessor are investigated in this section. The fixed priority model is important to consider for three main reasons. First, the fixed priority schema is one of the most simple and least time consuming to accomplish at run-time. Second, it is more robust than a dynamic priority model in that deadlines can be guaranteed for a certain set of the most important tasks even under transient overload. Finally, the fixed priority scheduling model is supported by Ada, the language mandated by the U.S. Department of Defense for all real-time and embedded programming applications.

The problem of scheduling periodic tasks with hard deadlines equal to the task periods on a uniprocessor system was first studied by Liu and Layland [11]. In their classic paper, they determined that rate monotonic scheduling was an optimal static priority scheduling algorithm. In rate monotonic scheduling, tasks are assigned priorities according to the length of their periods: the shorter the period, the higher the priority. The Liu and Layland bound, however, was pessimistic. An exact schedulability test was derived by Lehoczky, Sha, and Ding [10] to analyze task sets in which the utilization exceeds the Liu and Layland bound. The results to this point,

however, were limited to independent tasks. If concurrent tasks are expected to work together to accomplish a system goal, inter-task communication and cooperation are essential.

The exact schedulability analysis algorithm presented in [10] was extended by Sha, et al [13,14] to include the rate monotonic scheduling of periodic tasks that communicate via shared critical sections. They did this by observing that a direct application of common communication primitives can lead to unbounded priority inversion. They then proved that priority inversion can be bounded through the use of priority inheritance protocols. These extensions have made rate monotonic scheduling theory much more usable in the design of real-time systems, however, they are limited to communication via shared critical sections and do not accommodate the concerns of synchronous message passing.

Avrunin, Wiliden, and Corbett have shown that the execution time of tasks which use synchronous communications can be bounded through the use of a constrained expression approach [3,6]. Using the constrained expression approach, a synchronizing task set is expressed as a finite state machine or regular expression. Then, a set of inequalities is generated which define the execution constraints of the machine. The inequalities represent necessary conditions for any trace to be a behavior of the system. Integer programming techniques are then used to find a solution for the inequalities which yields an upper bound on the number of occurrences of the different events in the system. *This information can then be used to put an upper bound on the total execution time.* However, the bounds are only possible if all tasks use synchronous communications. Additionally, task scheduling is not taken into consideration. Under these circumstances, unbounded priority inversion can still be a problem.

A method for dealing with synchronizing tasks using earliest deadline scheduling has been presented by Mok [12]. Giering and Baker extended Mok's techniques to deterministic scheduling of a more general Ada tasking model with server tasks that can have select statements [8]. The earliest deadline scheduling method, however, is not currently supported by Ada 83, and this is not expected to change with Ada 9x [1,2].

The existing analysis and design techniques fall short when considering periodic tasks which communicate directly through the use of synchronous message passing as is found in the Ada rendezvous or blocking send and receive primitives in general. It is demonstrated in this section that the priority inheritance protocol can be extended to accommodate the scheduling of periodic tasks which communicate directly via synchronous message passing. The effects of synchronous message passing upon periodic tasks will be outlined, and sufficient conditions for schedulability of synchronizing periodic tasks will be given.

21.2 COMMUNICATION MODELS

21.2.1 Problem Description

Since inter-task communication and synchronization terminology varies, the underlying differences between the inter-task communication and cooperation models will be examined here. There are primarily two complementary inter-task communication and cooperation schemes: shared memory and message systems [16]. Using the shared memory system, tasks communicate, or exchange information, through the use of shared data structures. The message system scheme allows tasks to exchange messages directly. The shared memory scheme can also be described as a *resource sharing model*. Under the resource sharing model, processes obtain mutually exclusive access to shared resources. These resources may be shared memory, as in a critical section, region, or monitor, or a shared data bus or sensor. The physical resource is unimportant, the main consideration is that the use of the resource is mutually exclusive. For this reason, the terms shared memory and shared resources can be used interchangeably. A specific subclass of message systems which use synchronous message passing will be analyzed in this paper. The term *synchronization model* will be used to define this subclass. These two communication models are not mutually exclusive, and could be used simultaneously within a single system.

To use shared memory communications, a common variable or data structure is usually guarded by a semaphore or enclosed in a monitor. The main purpose for guarding a critical region with a semaphore is to preserve data consistency by ensuring mutually exclusive access to the shared data. The decision as to whether a task is allowed to enter a critical region is contingent only upon the current state of the system, i.e. whether another task is currently executing within the critical region, and not on the past activities of other tasks. Critical regions guarded by a semaphore can be modeled with a *passive* Ada monitor task as described in [13]. However, even though periodic tasks make entry calls to (or synchronize with) the monitor task, this implementation of a critical region should not be confused with direct synchronization of periodic tasks because the tasks do not call each other directly. In fact, a restriction on direct synchronization is explicitly stated in [13]. Tasks sharing the data accessed in the critical region are dependent because they share data, but order-independent because a sequencing of events is not imposed.

The resource sharing model can lead to unbounded priority inversion. Unbounded priority inversion is the situation where a higher priority task is blocked by a lower priority task for an undetermined amount of time. The classical example given has three tasks τ_1 , τ_2 , and τ_3 in descending order of priority with τ_1 holding the highest priority. τ_1 and τ_3 share a common data structure D which is accessed only in a critical section guarded by a binary semaphore. It could be the case that τ_3 enters the critical section first, and then is preempted by the higher priority task τ_1 . τ_1 could then attempt to enter the critical section, but be blocked by the lower priority task τ_3 which

already has a lock on the semaphore. This situation is called priority inversion because the high priority task τ_1 is being blocked by the low priority task τ_3 . The priority inversion is unbounded because when τ_1 is blocked by τ_3 , the intermediate priority task τ_2 could then preempt τ_3 and execute (blocking τ_1) as long as it wants. In fact, the blocking time is indeterminate because any other intermediate priority tasks in the system could also execute while τ_3 is blocking τ_1 .

A major disadvantage with using shared memory communications is that data may be over-written before another task has a chance to read it. If it is critical that no data is lost due to over-writes, then solutions could include bounded buffers, monitors, or synchronous message passing. Additionally, the system may have a requirement for task sequencing. However, the use of these solutions can impose an ordering on the tasks which can invalidate the priority inheritance approach. Two major assumptions for the use of the priority inheritance approach in [14] are:

1. "A job is a sequence of instructions that will continuously use the processor until its completion if it is executing alone on the processor."
2. "The critical sections of a job are properly nested and a job will release all of its locks. if it holds any, before or at the end of its execution."

It is well known that bounded buffers, monitors, and synchronous message passing can be implemented with semaphores, however, any job that uses any of the implementations will violate the two assumptions for the priority inheritance approach stated above. Any task that uses semaphores for sequencing will violate assumption (1) because the task's completion may depend upon the execution of another task. Furthermore, when semaphores are used to enforce a sequence between two tasks, one task usually locks a semaphore, while the other task unlocks the semaphore; thus violating assumption (2). For these reasons, another approach is needed when analyzing tasks which use sequencing primitives.

Unbounded priority inversion is still a problem under the synchronization model. Consider a situation similar to the unbounded priority inversion problem described previously. The system is comprised of three tasks τ_1 , τ_2 , and τ_3 in descending order of priority with τ_1 holding the highest priority. τ_1 and τ_3 engage in synchronous communications, while τ_2 does not. It could be the case that τ_1 arrives at the synchronization point first, and then is blocked while waiting for τ_3 to execute up to its corresponding synchronization point. This is an example of priority inversion because the high priority task τ_1 is being blocked by the low priority task τ_3 . The priority inversion is unbounded because while τ_1 is blocked by τ_3 , the intermediate priority task τ_2 (and any other intermediate priority tasks) could then preempt τ_3 and execute for an indeterminate amount of time. Therefore, unless all tasks synchronize, an execution bound for a task set cannot be determined without also considering scheduling disciplines. In the remainder of this section, it

will be shown that priority inheritance can be extended to the synchronization model and an algorithm for conducting schedulability analysis of synchronizing task sets will be offered.

21.2.2 Synchronization Model

Almost every concurrent language or distributed system has some form of synchronous message passing available. The blocking send/receive primitives and the Ada rendezvous are common examples of synchronous message passing. This paper will consider the schedulability implications of concurrent tasks which use named synchronous communications. This communication model is called symmetric because both the sender and the receiver must specify the destination or source of the message, respectively. To help convey the idea that the analysis techniques proposed in this paper are language independent, the generic term SYNC will be used to describe an occurrence of synchronous message passing. SYNC has the following syntax:

"SYNC message to/from task".

Consider, as an example, two tasks τ_1 and τ_2 that need to transfer a message X. The SYNC maps directly to the named blocking send and receive primitives as follows:

τ_1 : "send X to τ_2 " τ_2 : "receive X from τ_1 "

If the two statements shown above match up in an execution of a concurrent program so that the message X is transferred, the two statements are defined as a *synchronization pair*, or just *SYNC-pair*. The SYNC does not map directly to the Ada rendezvous because the rendezvous is asymmetric in that receivers cannot name the sender. To use the techniques presented here, it is necessary to enforce a programming discipline in which each calling task has its own entry in the called task. To aid in analysis, the name of each entry can be prefixed by its caller's name as shown below:

τ_1 : " $\tau_2.\tau_1$ message(X)" τ_2 : "accept τ_1 message(X: in ITEM)"

For the restricted form of the Ada rendezvous to work as desired, only τ_1 must call $\tau_2.\tau_1$ message. This paper adopts the same definition of *job* as is found in [14]. The basic definition is restated here:

A *job* is a sequence of instructions that will continuously use the processor until its completion if it is executing alone on the processor; that is, assume that jobs do not suspend themselves, say for I/O operations. However, such a situation can be accommodated by defining two or more jobs. In addition, assume that the critical sections of a job are properly nested and a job will release all of its locks, if it holds any, before or at the end of its execution.

Sha, et al [14] use the terms *job* and *task* interchangeably, and define a periodic task as a sequence of the same type of job occurring at regular intervals. This definition of a *task* is will be extended as follows to accommodate synchronous message passing:

A *task* is a sequence of jobs which may have SYNC events interleaved between them. A precise definition for a task is given by the regular expression $(J | \text{SYNC})^+$ where J denotes any job and SYNC denotes any SYNC. For example, τ_i might be specified by $J_{i1}, J_{i2}, \text{SYNC}_{i1}, J_{i3}, \text{SYNC}_{i2}$. A *periodic task* is a task that is dispatched at regular (periodic) intervals.

For notational convenience tasks will be describe by τ_1, τ_2 , etc. with periods of T_1, T_2 , etc. Unless otherwise specified, the priority of a given task is equal to its number with lower numbers indicating higher priorities. It is not necessary for the duration of jobs and SYNC's to be deterministic. However, it is assumed that an upper bound on the sequential execution times of each job and SYNC is known from a priori timing analysis using either language or runtime analysis techniques such as can be found in [5, 15, 17]. Also, it is assumed that the SYNC's are not conditional. The analysis presented in this paper is done under the assumption that the context switch time is zero. This restriction is easily relaxed by adding the context switch time to the execution time of the preempting task or SYNC event.

Each task has a fixed base priority. Because the order in which synchronizing tasks execute is important, tasks which participate in SYNC's are required to have unique priorities. Tasks that do not participate in SYNC's and have equal periods may be assigned the same priority. The non-synchronizing tasks with equal priorities can then be replaced by an equivalent single task for schedulability analysis purposes. Multiple tasks which run on a single processor are scheduled using a preemptive priority-driven discipline. It is always the case that the ready task with the highest priority is executed on the processor.

21.3 SCHEDULABILITY ANALYSIS OF SYNCHRONIZING TASKS

For Sections 21.3 through 21.5, it will be assumed that tasks only communicate through synchronous message passing, and do not use shared resources. This assumption simplifies the discussion and analysis. In Section 21.6, it will be shown that this assumption can be relaxed and that the techniques presented here can be used for task systems that use both synchronization and resource sharing models of communication.

21.3.1 Issues When Tasks Synchronize Directly

A number of problems are encountered when tasks synchronize directly. Synchronous message passing differs from shared memory communications because a partial ordering of events is enforced. Thus, the time required by the communication is contingent not only upon the time required to exchange the message, but also upon past execution of another task (or tasks).

The time necessary to complete an individual task is influenced by three things: its own execution time, time during which it is preempted by a higher priority task, and blocking time. The first two

times are addressed in Liu and Layland's original paper. The main sources of blocking examined by Sha, et al are the use of shared resources, and the execution of tasks which do not obey rate monotonic priorities (such as interrupts). Both types of blocking can lead to priority inversion. Uncontrolled priority inversion can lead to high priority tasks being blocked indirectly by lower priority tasks for indefinite periods of time. Forms of priority inheritance which can be used to bound priority inversion are described in [14].

Additional sources of blocking are found in a system of synchronizing tasks. First, a task which needs to SYNC with a lower priority task may be blocked while the lower priority task executes up to its synchronization point. As previously discussed, this case can lead to unbounded priority inversion if left uncontrolled. Second, a task can be delayed while waiting to SYNC with a higher priority task. This second type of blocking is encountered when the higher priority task has completed execution during its current period and is suspended until the start of its next period. Priority inheritance has been shown to be useful in bounding priority inversion due to the use of shared critical sections. The following extension of priority inheritance to directly synchronizing tasks is used here.

Definition: basic priority inheritance protocol for *directly synchronizing tasks*:

- When a task initiates a SYNC, the lower priority task of the pair inherits the priority of the higher priority task.
- The inheritance is transitive: If the task inheriting the priority described above is currently blocked for the same reason, the inheritance is propagated to the blocking task.
- Tasks return to normal priorities as SYNC's are completed.

The above defined priority inheritance protocol is consistent with the Ada 9x Mapping documents and the Catalogue of Interface Features and Options [1, 2, 4]. When basic priority inheritance is used for directly synchronizing tasks, the completion time for each task can be bounded. An algorithm for calculating the delays precipitated by synchronizations, and the resulting worst case task completion times will be given in the next section. Although priorities are no longer static when priority inheritance is used, the priority changes are controlled and can easily be carried out at runtime.

21.3.2 Task Execution Graph and Schedulability Analysis Algorithm

It has been shown that priority inheritance solves the unbounded priority inversion problem. When a higher priority task must wait for a lower priority task to execute up to a synchronization point, the lower priority task inherits the higher priority and, therefore, cannot be preempted by an intermediate priority task. To analyze the timing properties of synchronizing tasks, the following execution graph is introduced:

- A. Vertices represent the following events:
- Box - Start of task period
 - Circle - SYNC
 - Triangle - Task completion for current period.
- B. For two events A_i and B_i in task i , " A_i precedes B_i " is denoted by a uni-directional edge from A_i to B_i . Define this edge as an *execution edge*. An execution edge represents the execution of one or more jobs between events A_i and B_i . The duration of this edge is the worst case sequential execution time required by task i to complete these jobs.
- C. Each SYNC from task i to task j is denoted by a bi-directional edge between S_{ij} and S_{ji} . Define this edge as a *synchronization edge* and the duration of this edge is the time required to complete the synchronization in the worst case.
- D. Each edge is labeled with its duration, and the goal is to assign a worst case completion time to each node.

An example of an execution graph for a set of three non-periodic tasks is shown in Figure 1. The simple non-periodic example is given for the purpose of explaining the schedulability analysis algorithm. A more complicated periodic example will be given in Section 21.5.

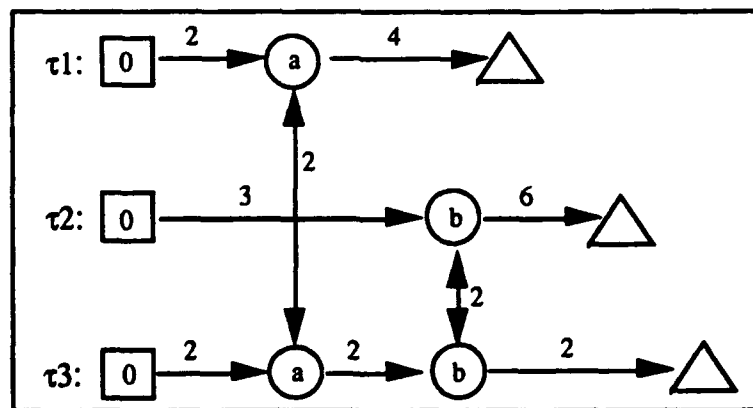


Figure 1: Example Execution Graph

The SYNC's enforce a partial ordering of events which is similar to Lamport's happened before relation [9]. All nodes and edges in the execution graph are events in the real system. The nodes represent completion times for particular events which are distinct points in time, and the edges are processes which take time to execute on the processor. Let all such events, not necessarily in the same task, be denoted by E_n . Without considering priority, the partial ordering can be described by the following rules:

1. if $E_1 \rightarrow E_2$ and $E_2 \rightarrow E_3$ then $E_1 \rightarrow E_3$. -- Precedence.
2. if $E_1 \rightarrow E_2$ and $E_2 \rightarrow E_1$ then $E_1 \leftrightarrow E_2$. -- Synchronization.
3. if $\text{not}(E_1 \rightarrow E_2)$ and $\text{not}(E_2 \rightarrow E_1)$ then $E_1 \parallel E_2$. -- Concurrency.

A result of Rule 1 is that if E_2 is reachable from E_1 in the execution graph, then E_1 precedes E_2 . Rule 2 will only apply to SYNC-pairs which must happen at the same time. $E_1 \parallel E_2$ means that E_1 and E_2 could take place concurrently. To provide a more precise definition of *synchronizing tasks*, it is said that two tasks *synchronize* if they engage in some form of synchronous message passing. And, a *synchronizing task set* is the set of all tasks related by the transitive closure of the above synchronization precedence relationship.

For all events that could potentially be executed concurrently, the order will be determined by priority. Furthermore, since all tasks have a unique priority, the events are totally ordered. The worst case completion time for each node can be calculated by summing the durations of each edge that precedes the node. When priority inheritance is used, a lower priority job could inherit the priority of a higher priority job. This inheritance would result in two jobs having the same priority. At first glance, this fact might seem to discredit the total ordering claim. However, since the inheritance can only take place if the higher priority job is blocked, the scheduler will never have to make the choice between two synchronizing jobs with the same priority. The highest priority task ready to run (τ_h) will always drive the execution. When priority inheritance is used, the completion times for the nodes in τ_h will only be effected by the events that lie on a path to the node or are in a higher priority task.

The description of the system as an execution graph, and the use of priority inheritance leads to a recursive graph algorithm which can be used to determine the completion times of each SYNC event and each task. A pseudo-code description of the algorithm called *schedule* can be found below. The input to the algorithm is a task execution graph and a system cycle. The task execution graph is as shown in Figure 1. Each task also has a pointer to keep track of the current node. For the example of a non-periodic task set, the system cycle is simply the deadline by which all tasks must complete. The algorithm is designed to traverse the graph in the order defined by priority and synchronizations and sum the durations attached to the execution and synchronization edges. The subprogram *execute* takes the task to analyze (*Current_Task*) and a target node (*Target*) as input and does the actual graph traversal. *Execute* starts traversing the graph from the current node pointer of *Current_Task*. *Target* is the node that determines how far *Execute* will traverse *Current_Task*. When *Execute* completes a node, it records the completion time and, if that node is an end node, it checks if the appropriate deadline was met. The base loop of the *schedule* algorithm runs until all tasks have completed, or the number of units of time specified by system cycle have been accounted for. In the non-periodic case, the highest priority task will have control of the analysis until it completes. In the periodic case, the highest priority ready task will have

control of the analysis until it completes, is blocked because its SYNC-partner is not ready, or a higher priority task becomes ready to run. A higher priority task will become ready to run at the start of its next period. The start of each task period is referred to as a scheduling point. If any of these termination events occur, then the *Execute* procedure will return, and the next highest priority task will be given control of the analysis. When the *Current_Task* arrives at a SYNC node, it recursively calls *Execute* on its SYNC-partner. When an edge is traversed (or *run*), time is incremented by the minimum of the edge duration and the time left to the next scheduling point of a higher priority task. If the whole duration of the edge is not recorded, then the partial run is marked, and the edge must be traversed again to account for the remainder of the time. If any deadline is missed before the system cycle is reached, then the task set is not schedulable, and the missed deadline is reported. If all deadlines are met, then the task set is pronounced schedulable.

An Ada program has been developed based upon the *Schedule* algorithm and its use as a schedulability analysis tool will be described with the example task set in Section 21.5.

Schedulability Analysis Algorithm

Global:

Task Execution Graph;
Task_Blocked, Time_Left : BOOLEAN;

procedure Schedule is

procedure Execute(Current_Task : Task_Id; Target:
Node_Type) is

Done : BOOLEAN := FALSE;

begin

while not Done and not Task_Blocked and
Time_Left loop

if the current node is not the target then

if the next edge is an execution edge then
run the next edge to completion or the next
scheduling point of a higher priority task;
if the edge was not run to completion then
Time_Left := FALSE;

end if;

else

-- The next edge must be a SYNC edge.

If the current node's sync-partner is ready then

-- The following recursive call simulates
-- priority inheritance.

Execute(sync-partner, SYNC w/ Current Node);

else

Task_Blocked := TRUE;

end if;

end if;

elsif the current node is an uncompleted
Sync_Type then

-- If this branch is taken, we know that we have

-- reached the target node which is a SYNC, and both
-- tasks are ready to SYNC.

run the SYNC to completion or the next
scheduling point of a higher priority task;

if the edge was not run to completion then

Time_Left := FALSE;

end if;

Done := Time_Left;

if Done then

Report the completion of the SYNC;

end if;

else

-- If this alternative is chosen, then we have

-- reached an End_Node which is the target.

Report the completion of the current task;

Done := True;

end if;

end loop;

if Task_Blocked then

-- All tasks on the call stack are blocked

elsif not Time_Left then

-- The start time for a higher priority task has been
-- reached.

end if;

end Execute;

begin

while Time < System Cycle and Tasks remain to be run

loop

Task_Blocked := FALSE;

Time_Left := TRUE;

Execute(Highest Priority Ready Task, End);

end loop;

end Schedule;

The solution to the graph from Figure 1 resulting from an application of the above algorithm is shown in Figure 2 below. The completed execution graph indicates that the SYNC labeled "a" will complete no later than time 6, the SYNC labeled by "b" will complete no later than time 17, and tasks 1, 2 and 3 will complete no later than times 10, 23 and 25 respectively.

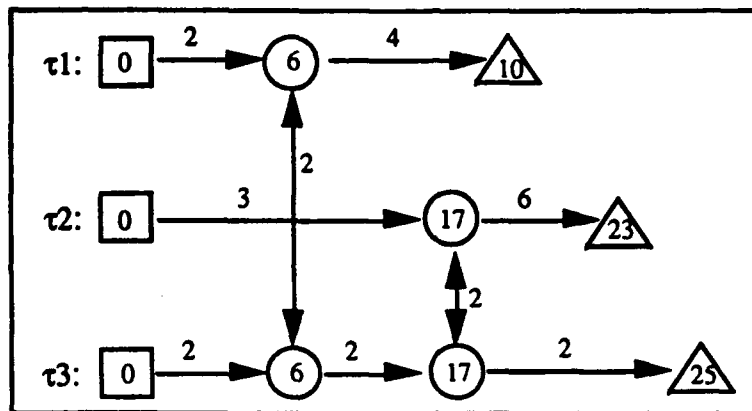


Figure 2: Solution to Figure 1

21.4 PERIODICITY OF SYNCHRONIZING TASKS

In the last section, an algorithm was presented to calculate worst case completion times for events in synchronizing task sets and determine whether those tasks meet deadlines. Another concern with directly synchronizing tasks is the effect of these synchronizations upon the period. A necessary relationship between the number of SYNC's and task periods is derived here. Also it will be shown that by analyzing a finite period of a synchronizing task set, the schedulability of the task set can be guaranteed.

21.4.1 Necessary Condition for the Periodicity of Synchronizing Tasks

An important issue that must be addressed when analyzing periodic tasks that synchronize is the effect of the SYNC's on the task periods. This problem can be described as a producer-consumer problem where one task (τ_1) produces synchronizations and another task (τ_2) consumes synchronizations. For the tasks to remain periodic, τ_1 and τ_2 must produce and consume SYNC's at the same rate. Otherwise, the task with the slower production/consumption rate will determine the maximum frequency. A relationship between synchronizations and task periods is described by the following theorem [7].

Theorem 1: For any two tasks τ_i and τ_j with periods T_i and T_j , let S_{ij} be the number of SYNC's in τ_i with τ_j . For a synchronizing task set with n tasks ($n \geq 2$), all tasks can meet periodic deadlines only if $S_{ij}/T_i = S_{ji}/T_j$ for every pair of tasks τ_i and τ_j in the set.

Note that this is a necessary condition only. The analysis described in section 21.3 is still required to guarantee that deadlines will be met. Tasks which do not synchronize can be initiated at any arbitrary rate as long as the utilization bound is satisfied, or an exact schedulability test is passed. The important point to note about synchronizing task sets is that the SYNC's impose a periodicity upon the system. Unless the tasks are initiated at periods which are consistent with the synchronization derived periodicity, the task set cannot meet its deadlines.

21.4.2 Determining Suitable Periods for Synchronizing Tasks

In this section, a technique will be outlined which uses the relationship described by Theorem 1 to calculate suitable periods for tasks in synchronizing task sets. Second, an iterative approach to calculate the minimal schedulable periods for a synchronizing task set with a given priority assignment will be described. Two $n \times n$ matrices S and T are used, where n is the number of dependent tasks in the set. Define S as the *synchronization matrix*, in which each entry s_{ij} represents the number of SYNC's from τ_i to τ_j . Define T as the *period matrix*, which is a diagonal matrix with each entry T_{ii} equal to the period of τ_i . ST is the matrix product of S and T . From Theorem 1, it is known that for the tasks to be periodic, the entries in ST must have the relationship $S_{ji} \cdot T_{ii} = S_{ij} \cdot T_{jj}$. In other words, ST must be a symmetric matrix for the tasks to be periodic. Given a task set in which the task periods are already known, a simple check for matrix symmetry will determine whether Theorem 1 holds for each pair of tasks.

If task periods are not given as part of the system specification, Theorem 1 can be used to synthesize suitable periods. From Theorem 1, it is known that each period in a synchronizing task set must be related to every other period in the set. To accomplish the synthesis, each period can be determined in terms of one period. Since it is known that the matrix ST must be symmetric, a linear equation can be created from each backward diagonal (i.e. each diagonal perpendicular to the main diagonal). Since all of the T_{ii} terms are unknown, there are n unknowns, and up to $2n-3$ equations - one for each backward diagonal. From Theorem 1, each equation is of the form:

$$S_{n,1}T_{1,1} + S_{n-1,2}T_{2,2} + \dots - S_{2,n-1}T_{n-1,n-1} - S_{1,n}T_{n,n} = 0 \quad (1)$$

Although the system can be overdetermined, the null solution of the form $T_{11} = T_{22} = \dots = T_{nn} = 0$, will always be a solution, so the system of equations will always be consistent. However, if the null solution is the only solution, the set of tasks is not schedulable with any period assignment because a period of 0 will not solve the physical problem. In the case of the null solution, the synchronization relationships of the task set must be modified before it can be schedulable.

Therefore, the coefficient matrix resulting from the equations described by (1) must have rank equal to $n-1$ for the task set to be schedulable. If this is the case, then a relationship for the periods can be established such that T_i must equal $b_i\alpha$ for some integer b and some number α . This relationship is called the synchronization derived period relationship (SDPR). A period assignment can only be chosen for the task set if it is consistent with the SDPR.

The following theorem establishes the viability of performing schedulability analysis for periodic synchronizing task sets [7].

Theorem 2: *Given a synchronizing task set containing n tasks, $\tau_1, \tau_2, \dots, \tau_n$ with periods T_1, T_2, \dots, T_n . Let the system cycle, denoted by C , be the smallest number such that $C = c_1T_1 = c_2T_2 = \dots = c_nT_n$ for integer constants c_1, c_2, \dots, c_n , and let all tasks start at time zero. For a given priority assignment, the synchronizing task set will always meet all of its deadlines if and only if it meets all of its deadlines during the first system cycle under worst case execution conditions.*

When the periods are integers, the system cycle will be the least common multiple of the periods. Define the LCM of the task periods as the *system cycle*. Theorem 2 means that if the schedulability analysis algorithm is used to verify that the task set meets all deadlines during the first system cycle, then all deadlines will always be met. Two forms of period synthesis can now be described. First, if a particular task period is required by the physical nature of the system being controlled, then given that period, the compatible periods for the rest of the tasks in the task set can be calculated from the SDPR. The schedulability of the given task set can then be determined by the schedulability analysis algorithm. The second form of synthesis uses the SDPR in an iterative approach for finding the minimum schedulable period assignment for a synchronizing task set. First, a lower bound on a schedulable period assignment can be established by setting the periods such that the processor utilization is 100%. Then, an upper bound can be established with any schedulable period assignment. Finally, the schedulability analysis algorithm can be used with an iterative interval splitting technique, such as is found in the binary search, to find the minimum schedulable period assignment to within any given degree of precision.

There are a few alternatives to be considered when it is determined that the task set is not schedulable. Some solutions might be:

1. Accept a later deadline.
2. Change the priority arrangement.
3. Change the code to reduce the worst case execution time.

Depending upon the physical system, Solution 1 may or may not be an option. In the examples given here, rate monotonic priority assignments are used. Although the rate monotonic priority

assignment strategy is usually reasonable, it is not necessarily optimal when tasks synchronize. Hence, different priority assignment may result in a schedulable system. For Solution 3, the schedulability analysis techniques presented can help narrow down the code that needs to be examined. First, if the SDPR does not hold, then nothing can be done to the execution times in the code to make the system schedulable. Second, since only events that precede an event can effect its completion time, only those execution and synchronization events that precede the missed deadline must be examined.

Finally, the execution graph can be used to conduct deadlock analysis. To conduct an a priori deadlock analysis using the execution graph, the following two points must be verified [18]:

- The SYNC statements correspond appropriately (i.e. send/receive or receive/send). This is referred to a single channel analysis.
- There must be no cycles (except between corresponding SYNC's).

Since the system of tasks is periodic, if no deadlock conditions exist in one system cycle, then the system is free from deadlock.

21.5 EXAMPLE SCHEDULABILITY ANALYSIS OF PERIODIC TASKS

In this section, an example will be given to demonstrate how the schedulability analysis algorithm described in Section 21.3 and the theorems from Section 21.4 can be combined in the design of a concurrent real-time computer program.

21.5.1 Sample Task Set

τ_1 : Sensor1:

```
loop
  delay until next start time
  execute          -- 2 units
  SYNC X to Controller -- 2 units
  execute          -- 4 units
end loop
```

τ_2 : Sensor2:

```
loop
  delay until next start time
  execute          -- 3 units
  SYNC Y to Controller -- 3 units
  execute          -- 6 units
end loop
```

τ_3 : Controller:

```
loop
  delay until next start time
  execute          -- 2 units
  SYNC X from Sensor1 -- 2 units
  execute          -- 3 units
  SYNC Y from Sensor2 -- 3 units
  execute          -- 3 units
  SYNC X from Sensor1 -- 2 units
  execute          -- 8 units
  SYNC Y from Sensor2 -- 3 units
  execute          -- 3 units
  SYNC X from Sensor1 -- 2 units
  control message to system -- 5 units
end loop
```

Figure 3: Sample Task Set

To illustrate the analysis techniques offered, consider the simple system shown in Figure 1 in which a controller task requires information from two sensors. In this example, the Controller task requires three readings from the Sensor1 task and two readings from the Sensor2 task to determine the appropriate control message each period. The worst case execution and synchronization times are shown in arbitrary units.

21.5.2 Determining Periods for the Sample Task Set

Suitable periods for the task set must first be determined. The period and synchronization matrices for the task set are:

$$S = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 3 & 2 & 0 \end{pmatrix} \quad T = \begin{pmatrix} T_1 & 0 & 0 \\ 0 & T_2 & 0 \\ 0 & 0 & T_3 \end{pmatrix}$$

The matrix product is:

$$ST = \begin{pmatrix} 0 & 0 & T_3 \\ 0 & 0 & T_3 \\ 3T_1 & 2T_2 & 0 \end{pmatrix}$$

Which yields the following set of equations to be solved:

$$\begin{aligned} 3T_1 & - 1T_3 = 0 \\ 2T_2 & - 1T_3 = 0 \end{aligned}$$

Since the system has a rank of 2, it has one free variable. Let $T_3 = \alpha$. Then $T_2 = \alpha/2$ and $T_1 = \alpha/3$. However, the periods must be integers, so the SDPR can be multiplied through by the least common multiple of the denominators giving the relationship $T_1 = 2\alpha$, $T_2 = 3\alpha$, and $T_3 = 6\alpha$. Given a periodic requirement for any one task in the system, the above relationship can be used to calculate the necessary periods for the remaining tasks. For example, if there was a requirement that the Controller task in Figure 1 execute once every 72 time units, the necessary periods for the two sensors of $T_1 = 24$ and $T_2 = 36$ could be calculated. Although this result may be apparent for the simple task set given, the solution may not be as apparent in a more complex task set with multiple task inter-dependencies.

21.5.3 Schedulability Analysis of the Sample Task Set

As was mentioned earlier, Theorem 1 represents a necessary condition for the periodicity of synchronizing tasks. A schedulability analysis is still required using the calculated periods. A system of periodic tasks will repeat after the least common multiple of its constituent task periods, or, the system cycle. If the task set meets all deadlines during the first system cycle, then all deadlines will always be met. From the results in section 21.5.2 for the task set given in Figure 1,

it can be determined that the system cycle time is $6\alpha = 72$. Therefore, the system cycle will contain three cycles of τ_1 , two cycles of τ_2 , and one cycle of τ_3 . An initial execution graph for the sample task set is shown in Figure 4 below.

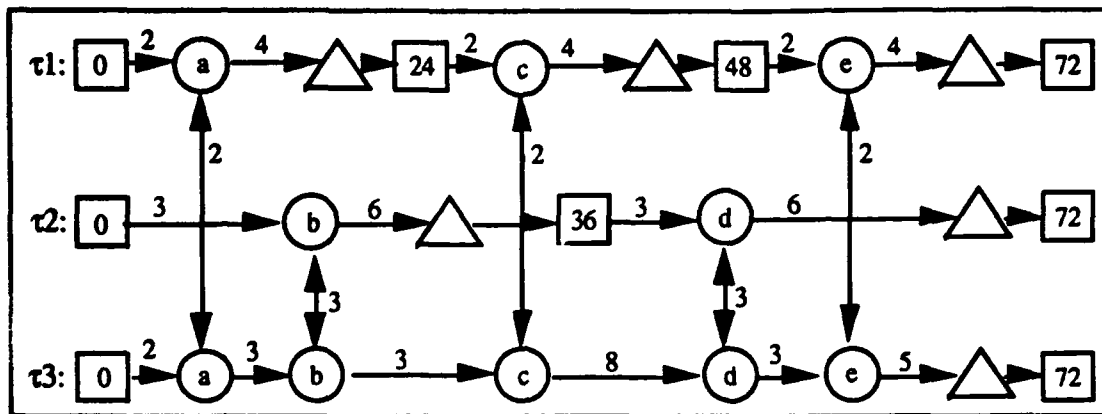


Figure 4: Initial Execution Graph

The above graph can now be used as input to the schedulability analysis program. The actual input is a specification of the task system as follows:

-- Example Task Specification

System_Cycle 72

-- Sensor1

Start_Task 1

Period 24

Execute 2

Sync to 3 2

Execute 4

End_Task 1

-- Sensor2

Start_Task 2

Period 36

Execute 3

Sync to 3 3

Execute 6

End_Task 2

-- Controller

Start_Task 3

Period 72

Execute 2

Sync from 1 2

Execute 3

Sync from 2 3

Execute 3

Sync from 1 2

Execute 8

Sync from 2 3

Execute 3

Sync from 1 2

Execute 5

End_Task 3

The schedulability analysis tool reports whether the system meets all deadlines and produces output that can be used to complete the synchronization graph. If the task set were not schedulable, then the schedulability analysis tool would report which task missed which deadline. It could also check if the deadline was missed due to a deadlock, or just because not enough time was available. The execution sequence experienced by the task set can be seen graphically in the Gantt chart depicted in Figure 5, and Figure 6 depicts the completed execution graph.

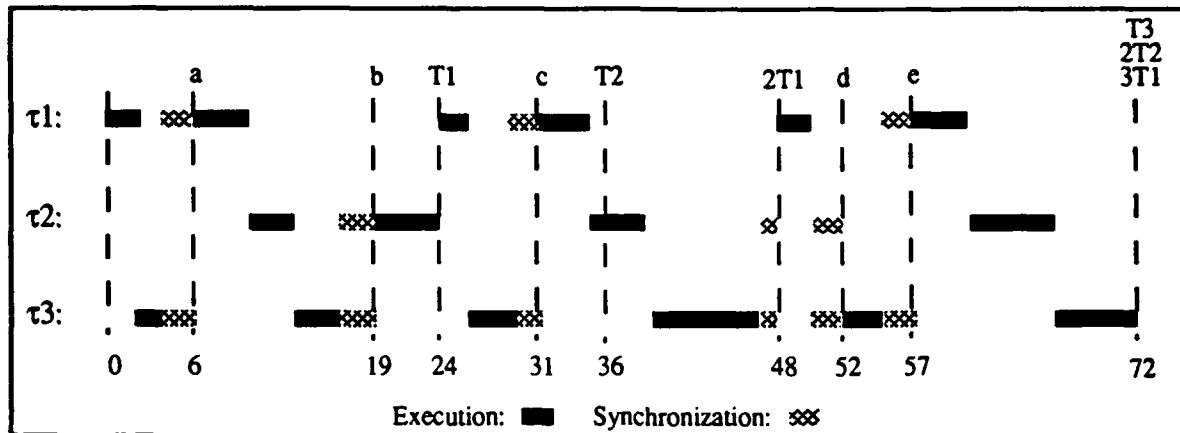


Figure 5: Gantt Chart for Given Task Set

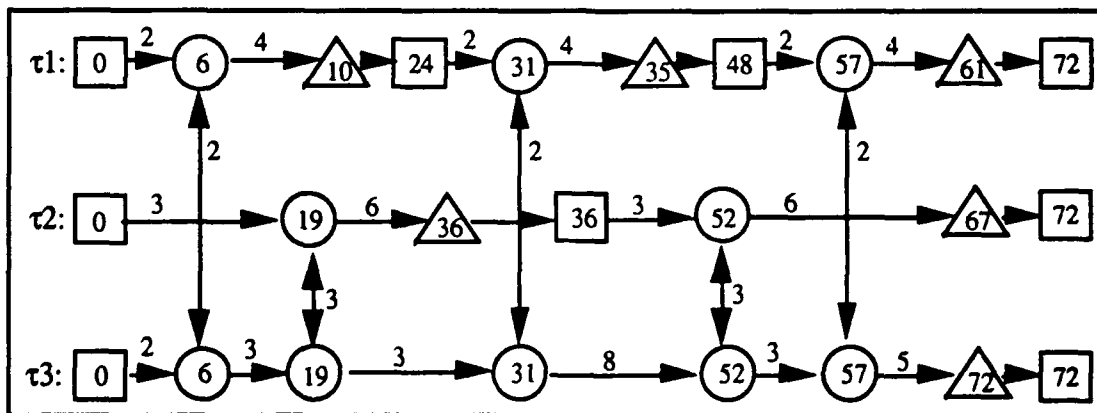


Figure 6: Completed Execution Graph

The analysis shows that all tasks meet all deadlines during the system cycle. Therefore, the task set is schedulable.

21.6 SCHEDULABILITY ANALYSIS WITH SHARED RESOURCES

In the last three sections, it has been shown that the delays due to direct synchronization are bounded when basic priority inheritance is used. Furthermore, an algorithm has been defined which can determine whether synchronizing periodic task sets with fixed base priorities are schedulable. However, the analysis technique presented so far has ignored the case when jobs access shared resources. By the definition adopted in Section 21.2.2, the jobs in synchronizing tasks may contain critical sections. In [7], it was shown that the upper bounds on blocking due to shared resources presented in [14] are still valid. This allows the algorithm to be used to determine whether all tasks can still meet their deadlines even if they are blocked for the worst case duration. This analysis can be accomplished by adding the worst case blocking time to the execution time for the given job and then running the schedulability analysis algorithm.

21.7 OTHER CONSIDERATIONS

21.7.1 Analysis of Ada Tasks

Ada tasks can be used as server tasks to simulate critical regions as described in [13]. When used for this purpose, it is appropriate for the tasks to use select statements.

However, to use the techniques presented here to make schedulability guarantees about periodic tasks which synchronize directly with the Ada rendezvous, all forms of non-determinism must be removed. The following restrictions on the Ada rendezvous accomplish this objective:

- Select statements must not be used.
- Entry points and entry calls must not be conditional.
- A separate entry point must exist for each individual task.

With the above restrictions, there can be no queueing of tasks on entry points, and all rendezvous will either be completed, or the initiating task will be blocked indefinitely.

21.7.2 Analysis of Tasks Which Have Conditional Synchronizations

The analysis presented here is based solely upon syntactical information. It is not possible to use syntactic analysis to make schedulability guarantees about task sets which use conditional synchronizations [7]. However, if it is possible to determine that the relationship holds for every semantically feasible path through the program, then the analysis can also be accomplished. It would then be necessary to conduct a schedulability analysis for every feasible path.

21.7.3 Time Units

The examples shown here use integer units of time. This is not a restriction for the use of the algorithm. However, computers generally deal with quantized units of time, such as the CPU cycle or timer interrupt rate. No matter how good it gets, the timing resolution available to the scheduler is limited. With these observations, it makes sense to deal with time as some integer multiple of the minimum time quantum.

21.8 CONCLUSIONS AND FUTURE WORK

This section has presented a methodology for analyzing directly synchronizing task sets. The priority inheritance protocol has been extended to accommodate directly synchronizing tasks and a schedulability analysis algorithm for calculating the completion times of synchronizing tasks has been provided. It has been shown that the bounds on priority inversion experienced due to the use of shared resources still apply to synchronizing task sets. It has also been shown that the periods of synchronizing tasks are related by their synchronization frequencies and this relationship results in synchronization derived periods (SDPR). An iterative approach for using the SDPR to determine a minimum schedulable system cycle has been given. Finally, the schedulability analysis algorithm provided has been implemented in an Ada program, and this program was used as a schedulability analysis tool for concurrent real-time programs that use both synchronization and resource sharing models.

Current work includes the development of schedulability analysis techniques for multiprocessor systems. Consider two typical multiprocessor systems.

System 1: Completely connected communication network, and one task per processor assignment. The schedulability analysis for this type system is almost trivial and can be done with a technique similar to the Critical Path Method of scheduling.

System 2: Bus inter-connection network, and one task per processor assignment. This system is more complicated than System 1. The bus can be treated as a shared resource. It is then possible to determine the maximum blocking that can occur due to mutually exclusive use of the bus, and then analyze System 2 as is done for System 1.

The goal of this research is the analysis of systems which allow multiple tasks on each of multiple processors, with all tasks being able to cooperate via both the synchronization model and the shared resource model. The algorithm presented in this paper uses a uni-processor interleaving of instructions. Namely, given two jobs J1 and J2 running on the same processor, the execution time for both jobs will be $\text{duration}(J1) + \text{duration}(J2)$. However, on a multiprocessor system with J1 and J2 on different processors, the execution time for both jobs will be $\max[\text{duration}(J1),$

duration(J2)]. Current plans include the extension of the schedulability analysis algorithm to accommodate multiprocessor interleaving, as well as the investigation of other schedulability analysis problems posed by multiprocessor systems.

Other future work includes the extension of these analysis techniques to a more general Ada rendezvous model and more general modes of message passing.

References

- [1] Ada 9X Mapping Volume I Mapping Rationale, Version 3.1, Intermetrics, Inc., Cambridge, MA, 23 August 1991.
- [2] Ada 9X Mapping Volume II Mapping Rationale, Version 3.1, Intermetrics, Inc., Cambridge, MA, 23 August 1991.
- [3] George S. Avrunin and Jack C. Wileden, "Automated Analysis of Concurrent and Real-Time Software." *Proc. ONR Third Annual Workshop on the Foundations of Real-Time Computing*, October 1990, pp. 305-322.
- [4] Catalogue of Interface Features and Options for the Ada Runtime Environment, a special issue of *Ada Letters*, Vol. XI, No. 8, Fall 1991 (II).
- [5] Russell M. Clapp, Louis Duchesneau, Richard A. Volz, Trevor N. Mudge, and Timothy Schultze, "Toward Real-time Performance Benchmarks For Ada," *Communications of the ACM*, v29, 760-778 (August 1986).
- [7] Andre N. Fredette and Robert J. Fornaro, "Schedulability Analysis of Real-Time Synchronizing Tasks." Submitted to *IEEE Transactions on Software Engineering*.
- [6] James C. Corbett, "Constrained Expression Analysis of Multi-processor Real-time Systems." Extended abstract, *Proc. ONR Fourth Annual Review and Workshop on the Foundations of Real-Time Computing*, November 1991.
- [8] E. W. Giering III, and T. P. Baker, "Toward the Deterministic Scheduling of Ada Tasks," *Proc., IEEE Real-Time Systems Symposium*, CS Press, Los Alamitos, California, 1989, pp. 31-40.
- [9] L. Lamport, "Time, Clocks, and the Ordering of Events in a Distributed System," *Communications of the ACM*, Volume 21, Number 7, (July 1978), pp. 558-565.

- [10] John Lehoczky, Lui Sha and Ye Ding, "The Rate Monotonic Scheduling Algorithm: Exact Characterization and Average Case Behavior." *Proc. IEEE Real-Time Systems Symposium*, CS Press, Los Alamitos, California, 1989, pp. 166-171.
- [11] C. L. Liu and J. W. Layland, "Scheduling Algorithms for Multiprogramming in a Hard Real Time Environment." *JACM* v20 (1) (1973), pp. 46-61
- [12] A. K. Mok, "Fundamental Design Problems of Distributed Systems for the Hard-Real-Time Environment," Ph.D. Thesis, MIT, May 1983.
- [13] Lui Sha and John B Goodenough, "Real-Time Scheduling Theory and Ada," *Computer*, Apr. 1990, pp. 53-62 .
- [14] Lui Sha, Rangunathan Rajkumar, and John P. Lehoczky, "Priority Inheritance Protocols: An Approach to Real-Time Synchronization," *IEEE Transactions on Computers*, Sep. 1990, pp. 1175-1185.
- [15] Alan C. Shaw, "Reasoning About Time in Higher-Level Language Software," *IEEE Transactions on Software Engineering*, Jul. 1989, pp. 875-889.
- [16] Abraham Silberschatz and James L. Peterson, *Operating System Concepts*, Addison-Wesley, 1988.
- [17] A. D. Stoyenko, V. C. Hamacher, R. C. Holt, "Analyzing Hard-Real-Time Programs for Guaranteed Schedulability," *IEEE Transactions on Software Engineering*, Aug. 1991, pp. 737-749.
- [18] K.C. Tai, "Sufficient Conditions for Freedom from Deadlocks in Distributed Programs," Extended Abstract, *Proc. ACM/ONR 1991 Workshop on Parallel and Distributed Debugging*.



**FACULTY, STAFF, AND STUDENTS
OF THE
PRECISION ENGINEERING CENTER**

FACULTY

THOMAS A. DOW

Director, Precision Engineering Center

Professor, Department of Mechanical and Aerospace Engineering

BS, Mechanical Engineering, Virginia Polytechnical Institute, 1966

MS, Engineering Design, Case Institute of Technology, 1968

PhD, Mechanical Engineering, Northwestern University, 1972

After receiving his PhD degree from Northwestern University in 1972, Dr. Dow joined the Tribology Section of Battelle Columbus Laboratories and worked there for ten years. His research interests were in the areas of friction and wear and included studies on a wide variety of topics from lubrication of cold-rolling mills using oil-in-water emulsions to wet braking effectiveness of bicycle brakes to elastohydrodynamic lubricant film generation in ball and roller bearings. He developed experimental apparatuses, established analytical models, and corroborated those analyses with experimental measurements. Dr. Dow joined the faculty at North Carolina State University in 1982 and was instrumental in developing the academic and research program in precision engineering. His current research interests include the design of precision machining systems, real-time control, and metrology. He was one of the founders of the American Society for Precision Engineering and currently acts as the Executive Director.

ROBERT J. FORNARO

Professor

Computer Science Department

BA, Mathematics, St. Vincent College, 1963

MA, Mathematics, The Pennsylvania State University, 1965

PhD, Computer Science, The Pennsylvania State University, 1969

Dr. Fornaro joined the Computer Science faculty at N.C. State in 1969. For the next several years he continued research in numerical analysis and contributed to the development of the Computer Science curriculum at NCSU. Since then he has been working in the computer systems area. He is especially interested in the relationship between operating systems software and multiprocessor architecture in real-time applications. He has developed an operating systems laboratory and an experimental concurrent programming environment in which high-level programming language constructs for concurrency and operating systems implementation can be compared and evaluated.

MICHAEL A. PAESLER

Professor

Department of Physics

BS, Physics and Math, Beloit College, Beloit, Wisconsin, 1968

MS, Physics, University of Chicago, 1971

PhD, Physics, University of Chicago, 1975

After receiving his graduate degrees at the University of Chicago, Dr. Paesler spent one year as a Visiting Scientist at the Max-Planck-Institut für Festkörperforschung in Stuttgart, FRG, working on the photo-transport in crystalline semiconductors. From 1977-80 he was a Research Fellow at Harvard University in the Division of Applied Sciences. His work there involved investigations of hydrogenated amorphous silicon. Since joining the North Carolina State University Physics Department faculty in 1980 and the Precision Engineering Center in 1985, Dr. Paesler has been involved in a number of investigations in both fundamental and applied physics and optics.

At North Carolina State University, Professor Paesler pursues research on several fronts. His work with the Precision Engineering Center focuses on the physics of material removal processes and the development of high resolution spectroscopies to monitor these processes. These efforts have centered on the non-destructive measurement of subsurface residual stress profiles of machined semiconductor surfaces, and super-resolution studies of such surfaces. In a closely related program, research is conducted on the development of optical microscopies and spectroscopies with sub-wavelength resolution. Emphasis in this program is primarily instrumental, with technique development the primary goal. Professor Paesler is also the director of the Research Experiences for Undergraduates program which has sponsored research for nearly 100 undergraduate students in Physics Department laboratories over the past five years.

PAUL I. RO

Assistant Professor

Mechanical and Aerospace Engineering Department

BS, Mechanical Engineering, University of Minnesota, 1982

MS, Mechanical Engineering, Massachusetts Institute of Technology, 1985

PhD Mechanical Engineering, Massachusetts Institute of Technology, 1988

From 1982 to 1988, Dr. Ro was a Research Assistant at the Laboratory of Manufacturing and Productivity at M.I.T., where he continued his research and gained expertise in various aspects of design and control of direct-drive robots. From 1986 to 1988, he was also a part-time engineering consultant at the Advanced Manufacturing Technology Division of Digital Equipment Corporation, where he dealt with various aspects of automation in tape drive assembly including vision inspection, design and implementation of a robotic workcell, and high-precision vision-based calibration.

In January 1989, Dr. Ro joined the faculty of North Carolina State University as an Assistant Professor in the Mechanical and Aerospace Engineering Department. Since then, he has been devoting his time teaching undergraduate and graduate level control courses. He has conducted research concentrating on characterizing and controlling micro-dynamic (small-input) behaviors of precision slide systems used for diamond turning, as well as developing other advanced control schemes such as Directional Dumping Control to improve surface quality of diamond-cut parts. His other research endeavor includes fixtureless assembly via two-arm coordination, marine and space robotics.

PHILLIP E. RUSSELL

Professor

Department of Materials Science and Engineering

BS Physics, Appalachian State University, 1975

MS Physics, West Virginia University, 1977

PhD Materials Science and Engineering, University of Florida, 1982

After graduate work at the University of Florida, Dr. Russell joined the Solar Energy Research Institute (a DOE lab) in Golden Co. in 1980. There he developed a photo-voltic materials and device characterization laboratory with emphasis on electron and ion beam analytical instrumentation. After three years at SERI, Dr. Russell joined JOEL, Inc. in Boston, Massachusetts, an electron optical instrumentation company where he led the technical and application groups. One of his major projects was the development of an electron beam based integrated circuit metrology system. He was also involved in the development and application of focused ion beam systems and electron beam lithography systems, as well as numerous analytical instrumentation projects.

On joining North Carolina State University, Dr. Russell took on the role of Director of the Analytical Instrumentation Facility and has established graduate level courses in electron optics and electron optical instrumentation techniques. He was awarded the NSF Presidential Young Investigator Award in 1987. His research at NCSU and the Precision Engineering Center are in the areas of Scanned Probe Microscopy, Focused Ion Beam Technology, Scanning Electron Microscopy, Lithography and beam testing of integrated circuits.

RONALD O. SCATTERGOOD

Professor

Materials Science and Engineering Department

BS Metallurgical Engineering, Lehigh University, 1961

MS Metallurgy, Massachusetts Institute of Technology, 1963

PhD Metallurgy, Massachusetts Institute of Technology, 1968

R.O. Scattergood is a Professor in the Department of Materials Science and Engineering. He received BS degrees in Mining Engineering and Metallurgical Engineering from Lehigh University. His MS and PhD degrees were obtained in Metallurgy from M.I.T. In 1968 he became a member of the basic research staff in the Materials Science Division at the Argonne National Laboratory. In 1981, he joined the faculty as a Professor of Materials Engineering at North Carolina State University.

Professor Scattergood's major research interests have been focused on the mechanical behavior of solids. He has worked in the areas of strengthening mechanisms in solids, continuum theory of defects, radiation and diffusion effects, wear and fracture processes in ceramics and precision engineering, with emphasis on fabrication processes. He has expertise in both analytical methods and computer modeling as well as in mechanical testing methods and microscopy. He has published over 110 technical and recently received the R.J. Reynolds Award for excellence in research and teaching.

JOHN S. STRENKOWSKI

Professor

Mechanical and Aerospace Engineering Department

BS, Aerospace Engineering, University of Virginia, 1972

**MS, Astronautics and Aeronautics,
Massachusetts Institute of Technology, 1973**

PhD, Applied Mechanics, University of Virginia, 1976

After completing his Ph.D., Dr. Strenkowski joined the Advanced Solid Mechanics Section of Battelle's Columbus Labs. While at Battelle, he participated in a variety of projects, which ranged from analyzing the vibrational and stress levels in offshore drilling platforms to predicting the residual stresses in weldments. This experience gave Dr. Strenkowski an awareness of the problems that face industry and the potential for their resolution through the application of advanced computational techniques.

Since joining the faculty of the Mechanical and Aerospace Engineering Department at North Carolina State University in 1978, Dr. Strenkowski has focused on computer-aided design, and finite element methods as applied to nonlinear problems involving large deformation and viscoplastic material behavior at elevated temperatures and high strain-rates. For the past eight years he has been developing new computational models of the cutting process. This work has brought new understanding to the mechanics of material removal processes, which will significantly change the way that cutting tools are designed to improve the productivity of cutting operations. Ultimately, this research is leading to new fundamental understanding that will result in significant advancements, such as the design of cutting tools for extended life and single-point diamond turning of ceramic materials.

STAFF

SALLY D. BIERCE

Secretary
Precision Engineering Center

Ms. Bierce was a guidance technician for the Wake County Public School System prior to joining the Precision Engineering Staff. This background provides experience in word processing, public relations, secretarial and other administrative responsibilities. Since joining the Center in October, 1989, Ms. Bierce has concentrated on mastering Latex on the MicroVax System, the Macintosh Computer and providing overall support for the Precision Engineering Faculty, Staff and Students.

KENNETH P. GARRARD

Research Assistant
Computer Science Department

BS, Computer Science, North Carolina State University, 1979
MS, Computer Studies, North Carolina State University, 1983

As a full-time research assistant, Mr. Garrard is studying the design of systems software that supports the development of high-speed real-time applications for special purpose multiprocessor computer systems. He has several years experience in academia and industry designing and implementing real-time systems. While in graduate school, Mr. Garrard's thesis work on concurrent programming environments was instrumental in the successful development of the Computer Science Department's Operating Systems Laboratory. He has also taught undergraduate courses in data structures, sorting, and systems programming. He has been employed as a scientific programmer for a pharmaceutical company, and as a consulting software engineer on a variety of projects, including real-time process control applications. As a Precision Engineering Center staff member, Mr. Garrard's current activities include the design and implementation of software for the Diamond Turning Machine Controller and the H²ART multiprocessor system.

GEORGE M. MOOREFIELD, II

**Research Assistant / Lecturer
Precision Engineering Center**

BS, Environmental Design / Architecture N.C. State University, 1977

BS, Mechanical Engineering, N.C. State University, 1983

Mr. Moorefield joined the Precision Engineering Center as a Research Assistant/Lecturer in February, 1989. Since joining, he has concentrated his efforts on creating a diamond turning facility which can produce prototype components for University and other interested parties, helping with laboratory courses related to the Center and assisting with the ongoing research agenda. He has an extensive background and interest in machining as well as equipment design and development. Prior to joining the Precision Engineering Center, Mr. Moorefield was employed as a Development Engineer in the New Process Engineering Department at AT&T Technologies, Inc.

LAUREN WILLIAM TAYLOR

**Research Assistant
Computer Science Department**

AS Electrical Engineering Technology, Behrend College, 1973

BS Computer Science, North Carolina State University, 1977

Mr. Taylor, a full-time research assistant, is involved in designing both hardware and software to support fast, real time multiprocessor computer systems. He has worked mainly with microprocessor based systems and was involved in setting up the Computer Science Department's Microprocessor Laboratory. He has taught undergraduate courses in programming concepts, computer architecture and microprocessor interfacing. He has also been employed as a systems analyst with an energy management company.

LEIGH ANN WEATHERS

**Administrative Assistant
Precision Engineering Center**

**BA Political Science, Appalachian State University, 1985
MPA Public Administration, North Carolina State University, 1990**

Ms. Weathers recently completed her Masters Degree in Public Administration at North Carolina State University with a minor in Economics. Prior to returning to graduate school, Ms. Weathers was a Branch Manager and Consumer Loan Officer with 1st Home Federal Savings and Loan Association in Cary, North Carolina. She brings to the Center skills in office administration, financial management, personnel management, and public relations.

LI ZHOU

**Postdoctoral Research Associate
Department of Materials Science and Engineering**

**BS, Physics, Peking University, 1987
PhD, Applied Physics, Oregon Graduate Institute, 1991**

Dr. Zhou joined the Precision Engineering Center as a Postdoctoral Research Associate in October, 1991. She received a PhD degree in Applied Physics from Oregon Graduate Institute. Her graduate research involved the investigation of the physics and properties of liquid metal ion sources (LMIS) and focused ion beams (FIB) modulated at high frequency by a focused laser beam and thermal effects. Currently Dr. Zhou has been conducting research programs on applications of electron microscopy and focused ion beam (FIB) techniques, especially on micro-fabrication of Atomic Force Microscope (AFM) tips and cross sectioning of integrated circuits by FIB micro-machining.

GRADUATE STUDENTS

JEFFREY A. ABLER is a PhD student in Mechanical Engineering. He received his MS in Mechanical Engineering from NCSU in 1991 and his BS from the University of Tennessee in 1989. The title of his master's thesis is "Control of Precision Slide Motion for Vibration Reduction in Diamond Turning".

WILLIAM D. ALLEN is a PhD student in Electrical & Computer Engineering. Mr. Allen received his BS and MS in Electrical Engineering from the University of Kentucky in 1965 and 1967 respectively. After graduating from UK, he worked for Harris Corporation and Simmonds Precision as an engineer and manager responsible for design of real-time computer control systems. He came to North Carolina State University in 1986 to pursue the PhD degree. His research interest is in multiprocessor computer architectures for real-time applications and is currently involved in studying the impact of interprocessor interactions on performance.

JAMES F. CUTTINO is a PhD student in Mechanical Engineering researching the mechanical aspects of a linear slide mechanism in order to achieve long range motion with nanometer accuracy. Having received his BS and MS degrees from Clemson University in 1985 and 1987 respectively, Mr. Cuttino worked for Michelin Americas Research and Development Corporation in Greenville, South Carolina for three years prior to arriving at NC State to continue his graduate studies.

JOSEPH D. DRESCHER is a PhD student in Mechanical Engineering studying the precision diamond point turning process. He is concentrating on development of a model for diamond turning which relates tool forces to surface quality.

ANDRE FREDETTE is a PhD student in Computer Science who is working on the development of structured design and analysis techniques for real-time computer systems. He is currently a Captain in the U.S. Army and is attending NCSU under the Army's Advanced Civil Schooling Program. He received the MS degree in Computer Science from NCSU in 1991 and the BS degree from the U.S. Military Academy at West Point in 1983.

GARY D. HIATT is a PhD student in Mechanical Engineering. Mr. Hiatt received his BS in Engineering Science & Mechanics from North Carolina State University and a MS in Engineering Mechanics from Virginia Polytechnic Institute & State University in 1981. After graduating from VPI, he worked as a mechanical engineer with the Naval Air Rework Facility at Cherry Point, North Carolina. Mr. Hiatt is working on developing a fracture mechanics model for single point diamond turning of brittle materials.

PETER I. HUBBEL received his MS in Electrical Engineering from NCSU in 1991. His thesis research involved developing control strategies for precision machine tool slides that exhibit non-linear dynamics. General interests include control theory and digital signal processing. He received his BA in Physics from Albion College in 1989.

WILLIAM C. LARSON is a PhD student in Mechanical Engineering researching the wear of diamond tools. After spending ten years as a nuclear submariner in the US Navy, and three years as an undergraduate instructor, he joined the Precision Engineering Center in 1989. Mr. Larson received a BS in Physics from Duke University (1976) and a ME in Nuclear Engineering (1983) from the University of Virginia.

MICHELE MILLER is a PhD student in Mechanical Engineering currently studying ductile regime grinding. She received her MS from NCSU in 1991 after doing DTM control research in the areas of fast tool implementation and geometric error correction. Upon receiving a BS from Duke University in 1986 she worked for General Motors for 1 1/2 years as a manufacturing engineer.

CHARLES B. MOONEY is a MS student in Materials Science and Engineering and is currently studying scanning probe microscopy tip technologies. He graduated from Western Carolina University in 1990 with a BS in Physics. Mr. Mooney was employed by Advanced Materials prior to his enrollment at North Carolina State University.

PATRICK J. MOYER is a PhD student in the Department of Physics. He is currently involved with developing an instrument used for performing high-resolution spectroscopic experiments. He received a BS degree (1986) and a MS degree (1988) in Physics from Moravian College and St. Bonaventure University, respectively.

G. WALTER ROSENBERGER is a MS student in Mechanical Engineering, investigating the parameters involved in grinding brittle materials. Mr. Rosenberger received his BS in Mechanical Engineering from Rose-Hulman Institute of Technology in 1985 and was a process engineer for Valmet Paper Machinery in Knoxville, TN prior to joining the Precision Engineering Center in 1991.

STANLEY M. SMITH received his BS degree in Chemical Engineering from North Carolina State University in 1986. He was a student member of AIChE. Mr. Smith received his MS in Materials Science and Engineering in December 1988 from NCSU. Under the auspices of the Thesis Parts Program Mr. Smith carried out his research at Argonne National Laboratory in the Materials and Components Technology Division. The title of his thesis is "Fabrication and Characterization of Al_2O_3/SiC -whisker Composites".

DWAYNE ALLEN SORRELL is a MS student in Computer Science. Mr. Sorrell received his BS in Computer Science from North Carolina State University in 1988. His research work is in the area of computer graphics for real-time application.

JOHN THORNTON is a MS student in Materials Science and Engineering who began working at the Precision Engineering Center in January, 1992. He graduated from Virginia Polytechnic Institute and State University in 1989 with a B.S. in Geology and an interest in microscopy. His research will be concentrated on the study of imaging mechanisms in the Atomic Force Microscope.

ROBERT M. TIDWELL received his MS degree in Materials Science and Engineering from NCSU in 1991. His research included diamond turning of brittle materials and image analysis of fracture damage on germanium. He received his BS degree in Physics from North Carolina State University in 1988.

GRADUATES OF THE PRECISION ENGINEERING CENTER

<u>Student</u>	<u>Degree</u>	<u>Date</u>	<u>Company/Location</u>
Jeffrey Abler	MS	December 1991	Pursuing PhD at NCSU Precision Engineering Ctr.
Kelly Allred	MS	June 88	
Tom Bifano	PhD	June 88	Boston University Boston, MA
Winston Scott Blackley	MS	May 1990	Mitsubishi Corporation Durham, NC
Peter Blake	PhD	December 88	NASA Goddard Greenbelt, MD
Mark Cagle	MS	June 86	NASA-Langley Norfolk, VA
John Carroll	PhD	January 86	Cummins Engine Co. Columbus, IN
Damon Christenbury	MS	June 85	Michelin Tire Co. Spartanburg, SC
William S. Enloe	MS	December 88	ITT Roanoke, VA
Karl Falter	MS	December 88	Pursuing PhD at NCSU in Mechanical Engineering
Steve Fawcett	PhD	July 91	Marshall Space Flight Ctr. M.S.F.C., Alabama
Jim Gleeson	MS	June 86	Battelle Columbus Labs Columbus, OH
David Grigg	MS	April 89	Pursuing PhD at NCSU in Material Science
Peter I. Hubbel	MS	December 91	Delco Electronics Kokomo, IN
Jerry Kannel	PhD	June 86	Battelle Columbus Labs Columbus, OH
Bryon K. Knight	MS	May 90	Harris Corporation Melbourne, FL

Mark Landy	MS	June 86	Battelle Columbus Labs Columbus, OH
Mike Loewenthal	MS	December 88	Cummins Engine Co. Columbus, IN
Michael Hung-Tai Luh	MS	June 89	Pursuing MS at Univ. of Cinn. School of Design
Michele Miller	MS	May 91	Pursuing PhD at NCSU Precision Engineering Ctr.
Gary Mitchum	MS	June 87	Harris Corporation Melbourne, FL
Larry Mosley	PhD	June 87	Intel Corporation Chandler, AZ
Hakan Ozisik	PhD	December 89	Aerospace Corporation Long Beach, CA
John Pellerin	MS	May 90	Pursuing PhD at Univ. of Kansas
Gordon Shedd	PhD	March 91	I.B.M. Research Laboratory Ruschlikon, Switzerland
Denise A. Skroch	MS	May 89	I.B.M. Corporation Raleigh, NC
Elizabeth F. Smith	MS	April 89	
Mary Beth Smith	MS	May 90	Harris Corporation Melbourne, FL
Ronald Sparks	PhD	May 91	Alcoa Corporation Pittsburg, PA
Michael Tidwell	MS	December 91	

ACADEMIC PROGRAM

Problems and limitations associated with precision manufacturing can originate in the machine, the process, or the material. In fact, most problems will probably be caused by a combination of these factors. Therefore, improvement of current processes and development of new manufacturing methods will require knowledge of a multi-disciplinary array of subjects. The educational goal of the Precision Engineering Center is to develop an academic program which will educate scientists and engineers in metrology, control, materials, and the manufacturing methods of precision engineering.

The graduate students involved in the Precision Engineering Center have an annual stipend as research assistants. They can take up to 3 classes each semester while spending about 20 hours per week on their research projects. These students will also work in the Center full-time during the summer months.

The Precision Engineering Center began in 1982 with an emphasis on the mechanical engineering problems associated with precision engineering. As a result, the original academic program proposed was biased toward courses related to mechanical design and analysis. However, as the research program has developed, the need for complementary research in sensors, materials, and computers has become obvious. A graduate student capable of making valuable contributions in the computer area, for example, will require a significantly different academic program than in mechanical engineering. For this reason, the Center faculty have set a core curriculum and each student in the program is required to take at least 3 of these core courses. The remainder of the courses for the MS or the PhD degree are determined by the university or department requirements and the faculty committee of the student.

The required courses are:

- MAE 589 Metrology and Precision Engineering
- PY 516 Physical Optics
- MAT 500 Modern Concepts in Materials Science
- CSE 574 Real Time Systems

PhD DEGREE PROGRAM

The PhD program in Precision Engineering has been set up as a multi-disciplinary program, drawing upon courses throughout the University to provide background and expertise for the students. It should contain required courses to insure solid grounding in the fundamentals plus electives to prepare the student in his area of specialization. Because Precision Engineering is concerned with an integrated manufacturing process, students interested in computer control, materials, machine structure, and measurement and actuation systems are involved in the program. Student research projects include the wide variety of topics addressed in this report. Each student's thesis should have an experimental component because Precision Engineering is basically a hands-on technology.

MS DEGREE PROGRAM

The Master of Science degree will have a higher percentage of application courses than the PhD degree. The emphasis will be to develop the foundation for involvement in precision engineering research and development. A total of 30 credits including 6 credits for the MS thesis is required. The thesis, while less comprehensive than the PhD dissertation, will be directed at important problems in Precision Engineering. Typically the MS program will take four semesters plus one summer.

UNDERGRADUATE PROGRAM

The undergraduate degree broadly prepares an engineering student for industrial activities ranging from product design and engineering sales to production implementation. Because a large share of engineers only have the BS degree, these will be the people who must implement the new technology developed in research programs like the Precision Engineering Center. Therefore, a way must be found to acquaint engineers at the BS level with the techniques, problems, and potential of precision manufacturing.

In most undergraduate degree programs only limited time is available for technical electives. However, these electives offer the student the opportunity to expand his knowledge in many different directions. Beginning graduate courses (such as metrology) can be used as undergraduate electives.

Undergraduate projects and summer employment have also been utilized to include undergraduate students into the research program of the Center. Two Material Science students were involved in the Center during the summer of 1988 under NSF funding directed by Professor Russell. In addition, two undergraduate students in Mechanical Engineering built a prototype laser interferometer as a senior project.

STUDY PLANS

Study plans for several example students are given below both for the MS and the PhD degree. Because of the breadth of the field and the wide range of thesis topics, few if any study plans will be exactly the same. The plan will depend upon the student's background, his interests, his thesis topic, the department, and the chairman and members of his committee.

PhD. PROGRAM IN MECHANICAL ENGINEERING

Major Courses:

- MAE 541 Advanced Machine Design I
- MAE 640 Advanced Machine Design II
- MAE 505 Heat Transfer Theory & Applications
- MAE 513 Vibrations of Mechanisms & Structural Components
- MAE 560 Computing Fluid Dynamics
- MAE 589 Metrology in Precision Engineering
- MAE 615 Nonlinear Vibrations
- MAE 619 Random Vibrations
- MAE 614 Mechanical Transfer & Machine Vibrations
- MAE 642 Machine Design Analysis
- MAE 699 Research

Minor Courses:

- MA 511 Advanced Calculus I
- MA 514 Methods of Applied Mathematics
- MA 530 Numerical Analysis II
- PY 516 Physical Optics
- ECE 516 System Control Engineering
- MAT 500 Modern Concepts in Materials Science
- ECE 613 Advanced Feedback Control
- ECE 555 Digital Image Processing

PhD. PROGRAM IN MATERIALS ENGINEERING

Major Courses:

- MAT 610 X-ray Diffraction
- MAT 699 Research
- MAT 500 Modern Concepts in Materials Science
- MAT 556 Composites
- MAT 615 Transmission Electronic Microscopy
- MAT 595b Defect Analysis
- MAT 633 Advanced Mechanical Props
- MAT 589 Scanning Electron Microscopy

Minor Courses:

- PY 414 Electricity & Magnetism
- STAT 515 Probability & Statistics
- MAE 541 Advanced Machine Design
- MAE 640 Advanced Machine Design II
- MAE 589 Metrology in Precision Engineering
- PY 516 Optics
- MA 401 Advanced Differential Equations

PhD. PROGRAM IN ME (FOR STUDENT WITH MS DEGREE)

- ECE 516 System Control Engineering
- ECE 591 Gate Array Design
- MAT 500 Modern Concepts in Materials Science
- PY 516 Physical Optics
- MA 502 Advanced Mathematics for Engineers
- MA 524 Math Methods in the Physical Sciences
- MA 530 Numerical Analysis II
- MAE 532 Fundamentals of Metal Machining Theory
- MAE 541 Advanced Machine Design I
- MAE 589 Metrology in Precision Engineering
- MAE 619 Random Vibrations
- MAE 640 Advanced Machine Design II

MS PROGRAM FOR ME STUDENT

- MAE 513 Vibrations of Mechanical & Structural Components
- MA 541 Advanced Machine Design I
- MAE 589 Metrology in Precision Engineering
- MAT 500 Modern Concepts in Material Science
- PY 516 Physical Optics
- MA 501 Advanced Math for Engineers and Scientists I
- MA 502 Advanced Math for Engineers and Scientists II
- MAE 699 Mechanical Engineering Research

MS PROGRAM FOR COMPUTER SCIENCE STUDENT

- CSE 501 Operating Systems
- CSE 506 Computer Architecture
- CSE 512 Compiler Construction
- ECE 520 Fundamentals of Logic Systems
- CSE 606 Concurrent Software Systems
- MAE 589S Metrology
- MAE 589W Digital Control Systems
- ECE 558 Digital Image Processing

MS PROGRAM FOR MATERIAL SCIENCE STUDENT

- MAT 500 Modern Concepts in Material Science
- MAT 510 Crystallography
- MAT 515 Transmission Electron Microscopy
- MAT 595 Scanning Electron Microscopy
- MAT 612 Advanced SEM
- MAE 589S Metrology
- PY 516 Physical Optics
- ECE 539 IC Technology and Fabrication
- MAT 699 Research

MS PROGRAM FOR PHYSICS STUDENT

- PY 516 Physical Optics
- PY 552 Introduction to Structure of Solids I
- PY 553 Introduction to Structure of Solids II
- PY 581 Quantum Mechanics I
- PY 582 Quantum Mechanics II
- PY 583 Advanced Classical Mechanics
- PY 585 Advanced Electricity and Magnetism I
- PY 586 Advanced Electricity and Magnetism II
- MAT 500 Modern Concepts in Material Science
- MAE 589S Metrology
- PY 699 Research

SHORT COURSES AND TV COURSES

Six graduate level courses, Scanning Electron Microscopy (MAT 512), Advanced SEM Surface Analysis (MAT 612), Modern Concepts in Material Science (MAT 500), Mechanical Properties of Materials (MAT 505), and Metrology (MAE 589) have been developed and taught as video courses offered nationwide via National Technological University. This past year approximately 120 students from industry and national laboratories participated in these courses. Future plans call for a MS program in Precision Engineering to be offered via the television network.

TECHNICAL REPORTS

Annual Report 1987	December 1987	336 pages
Semi Annual Progress Report	July 1988	24 pages
Annual Report 1988	December 1988	362 pages
Semi Annual Progress Report	September 1989	81 pages
Annual Report 1989	March 1990	357 pages
Semi Annual Progress Report	September 1990	119 pages

1986 PAPERS

Blake, P.N. and Scattergood, R. O., Chip Topography of Diamond-Turned Ductile Metals, SPIE 676, p. 96, August 1986.

Bryant, M.D. and Keltie, R.F., A Characterization of the Linear and Nonlinear Dynamic Performance of a Burleigh Piezoelectric Actuator - Part 1: Measurements, Sensors and Actuators, 9 (2), 95-104, 1986.

Carroll, J.T., Dow, T.A., and Strenkowski, J.S., Tool Force Measurement and Prediction in Diamond Turning, Proc. SPIE, Vol. 676, August 1986.

Russell, P.E., Materials Characterization, SEM Based Characterization Techniques, Ed. by N.W. Cleng, Marca Nicolet, published by Materials Research Society, Pittsburg, PA, 1986, ISBN 0-931837-35-9, (Book).

Strenkowski, J.S., Designing Tools with FEA, Machine Design, pp. 63-65, 1986.

THESIS/DISSERTATIONS

Cagle, C.M., Real-Time Control of Spindle Runout, M.S. Thesis, North Carolina State University, 1986.

Carroll, John Thomas III, A Numerical and Experimental Study of Single Point Diamond Machining, PhD. Dissertation, North Carolina State University, Raleigh, NC, 1986.

Gleeson, J.B., The Development of a Precision Translation Stage for the Study of High Speed Linear Positioning, M.S. Thesis, Department of Mechanical and Aerospace Engineering, NCSU, May 1986.

Landy, M.S., Modeling and Feedback Control of a Piezoelectrically Actuated Mechanical System, M.S. Thesis, North Carolina State University, 1986. ARTICLES Bifano, T., Dow, T.A., and Scattergood, R.O., Ductile Regime Grinding of Hard Materials, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 206, December 1986.

Blake, P., and Scattergood, R.O., Diamond Turning of Germanium and Silicon, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 180, December 1986.

Carroll, J.T., and Strenkowski, J.S., Tool Force Measurement and Prediction in Diamond Turning, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 107, December 1986.

Carroll, J.T. III, and Strenkowski, J.S., An Orthogonal Metal Cutting Model Based on an Eulerian Finite Element Method, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 92, December 1986.

Falter, P.J., The Development of a Fast Low Amplitude Tool Servo (FLATS), Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 164, December 1986.

Falter, P.J., and Dow, T.A., PAUL - Parallel Axis Ultraprecision Lathe, Third Annual Report on Precision Engineering - SRO 154, North Carolina State University, Raleigh, NC, p. 163, January 1986.

Fawcett, S.C. and Keltie, R.F., Amplitude Detector for Non-Contact Vibration Measurements, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 24, 1986.

Fornaro, R.J., Garrard, K., and Taylor, L., Systems Software for an 8086 Based Multiprocessor, TR-86-21, Computer Science Department Technical Report, North Carolina State University, Raleigh, NC., 1986.

Fornaro, R.J., Garrard, K., and Taylor, L., The Architecture of an 8086 Based Multiprocessor, TR-86-22, Computer Science Department Technical Report, North Carolina State University, Raleigh, NC., 1986.

Fornaro, R.J., Garrard, K., Taylor, L., A Structured Operating Systems Approach to High Speed Real-Time Control, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 48, December 1986.

Keltie, R.F., and Allred, C.K., Measurement of Structural Power Flow, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, pp. 30-41, December 1986.

Lehrman, S.A., and Strenkowski, J.S., Characterization of Material Behavior Using Torsion Testing, Orthogonal Metal Cutting and Finite Element Analysis, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 146, December 1986.

Loewenthal, M., and Dow, T.A., Basic Principles of CNC and DC Servo Amplifier Motor Control System, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 82, December 1986.

Luttrell, D.E. and Dow, T.A., Development of a High Speed System to Control Dynamic Behavior of Mechanical Structures, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 59, December 1986.

Mitchum, G.L., and Strenkowski, J.S., A Technique for Predicting Chip Separation in Orthogonal Metal Cutting, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 126, December 1986.

Ozisik, H., and Keltie, R.F., Structural Response Synthesis Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. IV, p. 42, December 1986.

Shedd, G. and Russell, P.E., Designing a Scanning Tunneling Microscope (STM) for the Characterization of Precisely Machined Surfaces, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol. 4, p. 8, 1986.

Smith, Elizabeth F. and Scattergood, Ronald O., Diamond Turning of Optical Glass, Precision Engineering Annual Report, North Carolina State University, Raleigh, NC., Vol IV, pp. 202-217, 1986.

PUBLICATIONS AND PRESENTATIONS

Papers published in Referred Journals

Development of a High-Speed System to Control the Dynamic Behavior of Mechanical Structures, D. E. Luttrell, T. A. Dow, Precision Engineering, October, 1987.

Design and Performance of a laboratory-scale Diamond Turning Machine, P. J. Falter T. A. Dow, Precision Engineering, October, 1987.

Characterization of the Linear and Nonlinear Dynamic Performance of a Burleigh Piezoelectric Actuator: Part 1 R. F. Keltie and M. D. Bryant, Sensors and Actuators, 9(2), pp. 95-104, 1986.

Chip Topography of Diamond-Turned Ductile Metals, P. N. Blake and R. O. Scattergood, SPIE 676, p. 96.

Precision Finishing of Ceramics, T. Bifano, P. Blake, T. Dow, and R. O. Scattergood, Proc. Symp. on Machining of Advanced Ceramics and Components, R. E. Barks, et. al., eds., ACS-ASME (1987).

Machining of Ceramics, T. Bifano, P. Blake, T. Dow, and R. O. Scattergood, Proc. 4th Int. Symp. on Opt. and Elec. App. Sci. Engr. The Hague, The Netherlands, in press (1987).

An Improved Finite Element Model of Orthogonal Metal Cutting, J. S. Strenkowski, NAMRC XV Conference Proceedings, 1987, pp. 506-509.

An Eulerian Finite Element Model of the Cutting Process, J. S. Strenkowski, submitted to International Journal of Mechanical Sciences, 1986.

Technical Reports

Annual Report 1986, January 1987, 255 pages.

Semi Annual Progress Report, July 1987, 24 pages.

Systems Software for an 8086 Based Multiprocessor, Fornaro, R. J., Garrard, K., and Taylor, L. TR-86-21.

The Architecture of an 8086 Based Multiprocessor, Fornaro, R. J., Garrard, K., and Taylor, L. TR-86-22.

Programmer's Manual for a Real-Time Heterogeneous Multiprocessor, Garrard, K., Fornaro, R. J., Taylor, L. W., and Skroch, D.

Further Development for Finite Element Model of Orthogonal Metal Cutting, Strenkowski, J. S., Final Report to the Lawrence Livermore National Lab, October, 1987.

Presentations

Precision Finishing of Ceramics, P. Blake, ACS Meeting, April (1987).

Development of a High-Speed System to Control the Dynamic Behavior of Mechanical Structures, D. A. Luttrell. ASPE Annual Meeting, Nov(1987).

Design and Performance of a laboratory-scale Diamond Turning Machine, P. Falter. ASPE Annual Meeting, Nov(1987).

Precision Engineering - Eastern Carolina Section of American Nuclear Society, January 30, 1987, T. A. Dow.

Precision Engineering, ONR-URI Meeting, Washington, DC, June 17-18, 1987, T. A. Dow.

Precision Engineering - Philips Research Laboratories, Eindhoven, The Netherlands, March 30, 1987, T. A. Dow.

Precision Engineering - Professional Engineers of North Carolina, February 6, 1987, T. A. Dow.

Precision Engineering - Sheffield Measurement Company, Dayton, Ohio, February 10, 1987, T. A. Dow.

Precision Finishing of Ceramics - SPIE Meeting, Micromachining of Elements with Optical and Other Submicrom Dimensional and Surface Specifications, The Hague, Holland, April 1-3, 1987, T. A. Dow.

Scanning Tunnelling Microscopy - Invited Lecture at the Carolinas Chapter of American Society for Metals, Raleigh, NC, April 1987, P. E. Russell.

Scanning Tunnelling Microscopy. A potential new technique for in situ Surface Chemistry Analysis at "Workshop on Plasma-Surface Interaction and Processing, of Materials, sponsored by Materials Research Society, RTP, NC, April 1987, P. E. Russell.

PUBLICATIONS

Papers

1. Bifano, T.F., Dow, T.A., and Scattergood, R.O., *Ductile-Regime Grinding of Brittle Materials: Experimental Results and the Development of a Model*, Proceedings of Advances in Optical Fabrication and Metrology Including Large Optics, SPIE Vol. 966, August 1988.
2. Bifano, T.G., Dow, T.A., and Scattergood, R.O., *Ductile-Regime Grinding: A New Technology for Machining Brittle Materials*, to be published ASME Transactions-Journal of Engineering in Industry, 1988. 8-18-88.
3. Bifano, T.G., Dow, T.A., and Scattergood, R.O., *Ductile-Regime Grinding of Brittle Materials*, Ultraprecision in Manufacturing Engineering, edited by Manfred Weck, and Robert Hartel, published by Springer-Verlag, p. 22, May 1988.
4. Blake, P., Bifano, T., Dow, T.A., and Scattergood, R.O., *Precision Machining of Ceramic Materials*, American Ceramic Society Bulletin, Vol. 67, No. 6, p. 1038, June 1988.
5. Blake, P.B. and Scattergood, R.O., *Ductile-Regime Turning of Germanium and Silicon*, to be published, Proc. of Symposium on Machining of Ceramic Materials and Components, ACS-ASME, December 1988.
6. Carroll, J.T., III and Strenkowski, J.S., *Finite Element Models of Orthogonal Cutting with Application to Single Point Diamond Turning*, to be published, International Journal of Mechanical Sciences, 1988.
7. Falter, P.J. and Dow, T.A., *A Diamond Turning Apparatus for Fabrication on Non-Rotationally Symmetric Surfaces*, Ultraprecision in Manufacturing Engineering, edited by Manfred Weck, and Robert Hartel, published by Springer-Verlag, p. 187, May 1988.
8. Fawcett, S.C. and Keltie, R.F., *Use Of A Fiber Optic Displacement Probe As A Surface Finish Sensor*, to be published Sensors and Actuators, 1989.
9. Fornaro, R.J., and Dow, T.A., *A High-Performance Machine Tool Controller*, to be presented 23rd IEEE Conference, Pittsburg, PA, October 1988.
10. Grigg, D.A. and Russell, P.E., *Vibration Isolation and Mechanical Designs for Several STM's in Air and UHV*, Proc. STM Conference, Oxford, England, July 1988.
11. Hren, J.J., and Shedd, G., *Field Electron Emission, The Atom Probe and Scanning Tunneling Spectroscopy*, Ultramicroscopy 24, p. 169, 1988.

12. Keltie, R.F. and Fawcett, S.C., *A Dual Channel Fiber Optic Displacement Probe for Structural Power Flow Measurements*, to be published Sensors and Actuators, 1989. 9-21-88.
13. Keltie, R.F. and Allred, C.K., *Measurement of Structural Power Flow in Vibrating Systems*, Proc. of the 6th International Model Analysis Conference, Orlando, Florida, pp. 1675-1681, 1988.
14. Keltie, R.F., and Fawcett, S.C., *Characterization of Fiber Optic Displacement Sensors for Precision Measurements*, Proc. of the 6th International Model Analysis Conference, Orlando, Florida, 1988.
15. Luttrell, D.E. and Dow, T.A., *Control of Precise Positioning System with Cascaded Colinear Actuators*, Proc. of the 1988 American Control Conference, Atlanta, GA, June 15-17, 1988.
16. Russell, P.E., and Grigg, D.A., *Mechanical Design Considerations for the Scanning Tunneling Microscope*, presented ACS Conference, Toronto, Canada, June 1988.
17. Russell, P.E., Griffis, D.P., Shedd, G.M., and Woodward, W.S., *A New Design for STM Control and Spectroscopic Data Acquisition*, ACS Conference, Toronto, Canada, June 1988.
18. Scattergood, R.O., Srinivasan, S., Bifano, T., and Dow, T.A., *R-Curve Effects for Machining and Wear of Ceramics*, presented at 7th International Symposium on Ceramics, Bologna, Italy, December 1988.
19. Sparks, R.G. and Paesler, M.A., *Micro-Raman Analysis of Stress in Machined Silicon and Germanium*, to be published, J. Precision Engineering, 1988.

Thesis/Dissertations

20. Bifano, T.G., *Ductile-Regime Grinding of Brittle Materials*, PhD Dissertation, North Carolina State University, 1988.
21. Blake, P.N., *Ductile-Regime Diamond Turning of Germanium and Silicon*, PhD Dissertation, North Carolina State University, 1988.
22. Fawcett, S.C., *Analysis and Development of a Multi-Channel Fiber Optic Probe for Structural Vibration Measurement*, M.S. Thesis, North Carolina State University, 1988.
23. Loewenthal, M.D., *Elastic Emission Polishing*, M.S. Thesis, North Carolina State University, 1988.

PUBLICATIONS

Papers

1. Moyer, P.J., Jahncke, C.L., & M.A. Paesler, *Spatially Resolved Spectroscopic Materials Characterization in the Evanescent Field*, submitted article to Physics Letters for publication November, 1989.
2. Musselman, Inga H. and Phillip E. Russell, *Platinum/Iridium Tips with Controlled Geometry for Scanning Tunneling Microscopy*, submitted for publication to Journal Of Vacuum Science and Technology, December 9, 1989.
3. Pellerin, J.G., G.M. Shedd, D.P. Griffis, and P.E. Russell, *Characterization of Focused Ion Beam Micromachined Features*, published in *Journal of Vacuum Science and Technology, B.*, Nov/Dec, 1989 issue.
4. Keltie, R.F. and Fawcett, S.C., *A Dual Channel Fiber Optic Displacement Probe For Structural Power Flow Measurements*, published in *Sensors and Actuators*, 19, (1989) 311-325.
5. Musselman, I.H., P.A. Peterson, and P.E. Russell, *Fabrication of Tips with Controlled Geometry for Scanning Tunneling Microscopy*, published in *Proceedings of IPES/ASPE Conference*, Monterey, CA, September 1989. In press, Precision Engineering, 1990.
6. Blake, P.N. and R.O. Scattergood, *Precision Machining of Germanium and Silicon*, *Journal of the American Ceramic Society*, in press, 1990.
7. Shedd, G.M. and P.E. Russell, *A Walking, Double-Tube STM Designed for use with a Focused Beam*, poster presented at American Vacuum Society, 36th National Symposium & Topical Conference, October 24, 1989, Boston, MA.
8. Srinivasan, S., Sparks, R., Pfeiffer, G., Scattergood, R., and M. Paesler, *Low Temperature Treatment of Transformation-Toughened Mg-PSZ-Solid Particle Erosion Studies*, in press, Journal of American Ceramics Society, 1989.
9. Bifano, T.G., Dow.T.A., and Scattergood, R.O., *Ductile-Regime Grinding: A New Technology for Machining Brittle Materials*, submitted to ASME Journal of Engineering for Industry, 1989.
10. Blackley, W.S. and R.O. Scattergood, *Diamond Turning of Brittle Materials*, to be published in *Proceedings of IPES/ASPE Conference*, Monterey, CA, September 1989.

11. Dow, T.A., T.G. Bifano, and C.M. Cagle, *Spindle Error Compensation*, presented and published at 1989 International Machine Tool Research Forum, Chicago, ILL, August 1989.
12. Drescher, J.D. and Dow, T.A., *Tool Force Model Development for Diamond Turning*, To be published in *Precision Engineering*, Poster presented at ASPE/IPES, Monterey, CA, September 1989.
13. Fawcett, S.C. and Keltie, R.F., *Use Of A Fiber Optic Displacement Probe As A Surface Finish Sensor*, to be published *Sensors and Actuators*, 1989.
14. Fornaro, R.J. and Davis, E.W., *Analysis and Implementation of Hierarchical Real-Time Architectures*, presented at Euromicro Workshop on Real-Time, Italy, 1989.
15. Musselman, I.H., R.-T. Chen, and P.E. Russell, *Roughness Measurements of Non-linear Optical Polymer Films by Scanning Tunneling Microscopy*, presented at the Electron Microscopy Society of America, San Antonio, TX, August 1989.
16. Musselman, I.H., and Russell, P.E., *Platinum Thin Film Roughness Measurements by Scanning Tunneling Microscopy*, published *Proceedings of the Microbeam Analysis Society Annual Meeting*, Asheville, NC, July 1989.
17. Musselman, Inga Holl, *Scanning Tunneling Microscopy and Atomic Force Microscopy*, *Analytical Chemistry*, Spring 1989.
18. Pellerin, J.G., G.M. Shedd, D.P. Griffiths, and P.E. Russell, *Micro and Nanofabrication with a Combined Focused Ion Beam/Scanning Tunneling Microscope*, *Microbeam Analysis-1989*, San Francisco Press, Inc., San Francisco, CA, 1989.
19. Rhatigan, J.L., Johnson, R.R., and Dow, T.A., *An Experimental Study of Thermoelastic Effects in Scuffing Failure of Sliding Lubricated Contacts*, *ASME Journal of Tribology*, Vol 111, No 1, Jan 1989, p 23-28.
20. Russell, P.E. and I.H. Musselman, *Scanning Tunneling Microscopy of Polymers: A Status Report*, presented at the Electron Microscopy Society of America, San Antonio, TX, August 1989.
21. Russell, P.E., D.A. Grigg, I.H. Musselman and G.M. Shedd, *Applications of Scanning Tunneling Microscopy*, Seminar given in Tus. Alabama. Abstract published in *Journal of Electron Microscopy Technique*, April 1989.
22. Russell, P.E., *Design and Use of STM's for Precision Engineering*, Seminar presented at AT&T Bell Labs, Murray Hill, NJ, January 1989.
23. Sparks, R.G., Enloe, W.S. and Paesler, M.A., *Micro-Raman Applications in Precision Engineering*, poster presented at International ASPE Conference, September 1989, Monterey, CA.

24. Sparks, R.G., W.S. Enloe and M.A. Paesler, *Development of a High Axial Resolution Micro-Raman Technique for Studying the Effects of Machine Parameters in Machined Semiconductors*, *Proceedings of Micro-Beam Analysis Society*, Asheville, NC, July 1989.
25. Srinivasn, S., R.O. Scattergood, G. Pfeiffer, R.G. Sparks and M.A. Paesler, *Low-Temperature Treatment of Transformation Toughened Mg-PSZ-A Solid-Particle Erosion Study*, to be published in *Journal of American Ceramics Society (communications)*, 1989.
26. Strenkowski, J.S., *Prediction of Built-Up Edge Formation in Orthogonal Cutting of Aluminum*, *NAMRC XVII Conference Proceedings*, 95-102, May 1989.
27. Strenkowski, J.S., M.H. Luh, *Thermal Analysis of Orthogonal Cutting Using A Thermo-Viscoplastic Finite Element Model*, to be published in *Proceedings of ASME Winter Annual Meeting*, San Francisco, CA, December 1989.

Thesis/Dissertations

1. Drescher, Joseph Dean, *Tool Force Measurement in Diamond Turning*, Master Thesis, North Carolina State University, September, 1989.
2. Falter, Karl J., *Refinement of Experimental Procedure for Structural Power Flow Measurement*, Master Thesis, North Carolina State University, February 1989.
3. Grigg, David A., *Mechanical Design of a Scanning Tunneling Microscope for the Observation of Machines Surfaces*, Master Thesis, North Carolina State University, April 1989.
4. Luh, Michael Hung-Tai, *Thermal Analysis of Orthogonal cutting Using a Thermo-Viscoplastic Finite Element Model*, Master Thesis, North Carolina State University, June 1989.
5. Ozisik, Hakan, *Development and Implementation of an Open Loop Control Technique for High Speed Micropositioning in a Single Point Diamond Turning Process*, PhD Dissertation, North Carolina State University, October 1989.
6. Skroch, Denise A., *A Hierarchical Architecture for Real-Time*, Master Thesis, North Carolina State University, May 1989.
7. Smith, Elizabeth F., *Single-Point Diamond Turning of Amorphous Thermoplastic Polymers*, Master Thesis, North Carolina State University, April 1989.

PUBLICATIONS

Papers

Fawcett, Steven C., and Dow, Thomas A., *Precision Contour Grinding of Brittle Materials* submitted to *Precision Engineering*, December, 1990.

Shedd, Gordon M., and Russell, Phillip E., *The Effects of Low-Energy Ion Impacts on Graphite Observed by Scanning Tunneling Microscopy* submitted to *Journal of Vacuum Science and Technology A*, October, 1990.

Shedd, Gordon M., and Russell, Phillip E., *A Simple Model for the Electron-Density Superstructures Observed During Scanning Tunneling Microscopy of Perturbed Graphite Surfaces* submitted to *Physical Review Letters*, October, 1990.

Miller, Michelle H., Dow, Thomas A. and Falter, Peter J., *Application of a Fast Tool Servo for Diamond Turning of Non-Rotationally Symmetric Optical Surfaces* submitted to *Precision Engineering*, September, 1990.

Paesler, M. A., Moyer, P.J. and Johnson, C. E., *Analytical Photon Scanning Tunneling Microscopy* published in *Physical Review B*, August, 1990.

Musselman, I.H., Russell, P.E. , Chen, R.T., Jamieson, M.E., and Sawyer, L.C., *Correlative STM, FESEM, and TEM Studies of Fibrillar Structures in Liquid Crystalline Polymers* abstract presented and published in *Proceedings of the XIIth International Congress for Electron Microscopy*. San Francisco Press, Inc. 1990.

Falter, Peter J. and Dow, Thomas A., *Diamond Turning of Non-Rotationally Symmetric Surfaces* abstract presented at *ASPE 90' Conference*, September, 1990, Rochester, NY.

Ozisik Hakan, Bifano, Thomas G. and Dow, T.A., *Application of a Novel Two Step Digital Control Algorithm for Precision Actuation*, abstract presented at *ASPE 90' Conference*, September, 1990, Rochester, NY.

Fornaro, Robert J., Garrard, Kenneth, P., and Taylor, Lauren W., *Architectures and Algorithms for Computer Control of High Precision Machine Tools*, abstract presented at *ASPE 90' Conference*, September, 1990, Rochester, NY.

Blackley, Scott and Scattergood, Ronald O., *Mechanics of Material Removal in Diamond Turning*, abstract presented at ASPE 90' Conference, September, 1990, Rochester, NY.

Ro, Paul I. and Colby, Roy S., *Nonlinear Control of a Precision Slide with Coulomb Friction and Stiction*, abstract presented at ASPE 90' Conference, September, 1990, Rochester, NY.

Blackley, W.S. and R.O. Scattergood, *Ductile-Regime Processes in Diamond Turned Germanium Chips*, submitted to *Journal of Engineering for Industry*, June 1990.

Blackley, W.S. and R.O. Scattergood, *Crystal Orientation Dependence of Machining Damage -A Stress Model*, submitted to *Journal of American Ceramic Society*, June 1990.

Blackley, W.S. and R.O. Scattergood, *Ductile-Regime Machining Model for Diamond Turning of Brittle Materials*, submitted to *Precision Engineering*, June 1990.

Fawcett, Steven and Thomas A. Dow, *Computer Simulations of Precision Ground Surfaces*, extended abstract presented at *The Science of Optical Finishing Topical Meeting Conf. Proc.*, Optical Society of America, Monterey, CA, June 10-12, 1990.

Fawcett, Steven and Thomas A. Dow, *Extension of the 1-D Critical Depth of Cut Model for Brittle Materials to 3-D Contour Grinding* extended abstract to be presented at *Annual Meeting of the American Society for Precision Engineering Conf. Proc.*, Rochester, N.Y., September 23-28, 1990.

Shedd, Gordon M. and Phillip E. Russell, *The Scanning Tunneling Microscope as a Tool for Nanofabrication*, published in *Nanotechnology* 1, pgs. 67-80 March 22, 1990.

Shedd, Gordon M. and Phillip E. Russell, *Experiments in Nanomodification with the Scanning Tunneling Microscope*, abstract presented at *Seminar at UNC-Chapel Hill*, February 22, 1990.

Miller, Michele and T.A. Dow, *Implementation of a Fast Tool Servo on a Diamond Turning Machine*, abstract submitted for *ASPE Annual Meeting*, Rochester, NY, September 24-27, 1990.

Scattergood, R.O., *Mechanics of Material Removal in Single-Point Diamond Turning of Brittle and Amorphous Optical Materials*, abstract to be presented at *The Science of Optical Finishing Topical Meeting*, Monterey, CA, June 10-12, 1990.

Dow, T.A. R.O. Scattergood, Ductile/Brittle Transition and Development of Ductile Mode Grinding Technology, submitted to the Special Issue of the Journal of the Japan Society of Precision Engineering, January 16, 1990.

Pellerin, J.G., and P.E. Russell, Focused Ion Beam Micromachining of SI, GAAs, and InP, presented abstract to 34th International Symposium of Electron, Ion, and Photon Beams, January, 1990.

Fawcett, Steven C., Small Amplitude Vibration Compensation for Precision Diamond Turning, submitted Oct., 89 for publication in the Precision Engineering Journal, 1990. Published in Precision Eng. Journal, April 1990, Vol.12, No. 2.

Sparks, R.G., Enloe, W.S., Wellman, J., Pfeiffer, G. Agarwal, S.C., and M.A. Paesler, Micro-Raman Applications in Precision Engineering, submitted abstract to Department of Physics and Precision Engineering Center January, 1990.

Musselman, Inga H. and Phillip E. Russell, Platinum/Iridium Tips with Controlled Geometry for Scanning Tunneling Microscopy, published in Journal Of Vacuum Science and Technology, A 8 (4), July/August, 1990

Srinivasan, S., Sparks, R., Pfeiffer, G., Scattergood, R., and M. Paesler, Low Temperature Treatment of Transformation-Toughened Partially Stabilized Magnesia-Doped Zirconia - A Solid Particle Erosion Studies, published in, Journal of American Ceramics Society 73, May 1990, pgs. 1421-24.

Fawcett, S.C. and Keltie, R.F., Use Of A Fiber Optic Displacement Probe As A Surface Finish Sensor, published Sensors and Actuators, Vol. 24. No. 1, pg. 5-14, 1990.

Thesis/Dissertations

Pellerin, John , *Fundamental Aspects of Focused Ion Beam Micromachining*, Master Thesis, North Carolina State University, May, 1990. Has not been mailed to affiliates.

Smith, Mary Beth, *SIMPLE: A Multiprocessor Programming Environment for Real-Time Applications*, Master Thesis, North Carolina State University, May, 1990.

Blackley, Winston Scott, *Single Point Diamond Turning of Brittle Materials*, Master Thesis, North Carolina State University, May, 1990.

Knight, Byron Franklin, *Development of a Testbed for Precision Linear Motion*, Master Thesis, North Carolina State University, May, 1990.

PUBLICATIONS

Papers

1. Abler, Jeffrey A. and Ro, Paul I., *Development of Directional Damping Control for Vibration Reduction in Precision Slide Motion*, submitted to ASME Journal of Vibration and Acoustics, December 18, 1991.
2. Hubbel, Peter I. and Ro, Paul I., *Model Reference Adaptive Control of Dual-Mode Micro/Macro Dynamics of Ball Screws for Nanometer Motion*, submitted to the ASME Journal of Dynamic Systems, Measurement, and Control, December 18, 1991.
3. Larson, William, Chern, SHin-Yuh, and Strenkowski, John S., *Prediction of Diamond Tool Wear in Precision Machining*, presented at the poster session of the American Society of Mechanical Engineers Winter Annual Meeting in Atlanta, GA on December 5, 1991.
4. Fredette, Andre, Allen, William D. and Fornaro, Robert J., *Schedulability Analysis of Real-Time Synchronizing Tasks*, submitted to the 12th IEEE Real-Time Systems Symposium, December 3-6, 1991.
5. Abler, Jeffrey A. and Ro, Paul I., *Development of a New Vibration Reduction Control Scheme for Precision Slide Motion*, submitted for 1991 ASPE Annual Conference, Sante Fe, NM, October 13-18, 1991.
6. Smith, Stanley M. and Scattergood, Ronald O. Scattergood, *Short-Crack Toughness Determination for Brittle Materials*, submitted for 1991 ASPE Annual Conference, Sante Fe, NM, October 13-18, 1991.
7. Tidwell, Michael and Scattergood, R.O., *Analysis and Modeling of Diamond Turning of Brittle Materials*, submitted for 1991 ASPE Annual Conference, Sante Fe, NM, October 13-18, 1991.
8. Miller, Michelle H., Garrard, Kenneth P., and Dow, Thomas A., *Controller Design for a Modern Diamond Turning Machine*, submitted for 1991 ASPE Annual Conference, Sante Fe, NM, October 13-18, 1991.

9. Hubbel, Peter I. and Ro, Paul I., *Model Reference Adaptive Control of the Nonlinear Microdynamics of a Ball-Screw Driven Precision Slide*, submitted at 1991 ASPE Annual Conference, Sante Fe, October 13-18, 1991.
10. Cuttino, James F. and Knight, Byron F. and Dow, Thomas A., *The Preloaded Ball Screw As A Nanometer Motion Device*, submitted at 1991 ASPE Annual Conference, Sante Fe, October 13-18, 1991.
11. Mooney, C.B. and Russell, P.E., *Microscopy of Grown Tips for Tunneling Microscopy*, submitted to Tenth Anniversary Symposium on Advances in Microscopy, September 24-26, 1991.
12. Sparks, R.G. and M.A. Paesler, *"Depth Profiling of Residual Stress Along Tool Shoulders In Machined Germanium Crystals"*, submitted for publication to Journal Of Applied Physics, August 1991.
13. Ximen, Hong-Yu, and Phillip E. Russell, *Microfabrication of AFM Tips By Using Combination of Focused Ion and Electron Beam Techniques*, submitted at International Conference on Scanning Tunneling Microscopy, August 12-16, 1991.
14. Smith, Stanley M. and Scattergood, Ronald O., *Crack Shape Effects in Fracture Toughness Measurements*, presented at American Ceramic Society 93rd Annual Meeting & Exposition, April 1991, Cincinnati, OH, and submitted to Journal of American Ceramic Society, July, 1991.
15. Blackley, W.S. and R.O. Scattergood, *Ductile-Regime Processes in Diamond Turned Germanium Chips*, published in Precision Engineering, April 1991, Vol. 13, No. 2.
16. Allen, William D. and Fornaro, Robert J., *Application of Real-Time Scheduling Theory to Multiprocessor Pipelines* submitted for the Eighth Workshop on Real-Time Systems and Software (Atlanta), January, 1991.
17. Musselman, Inga H., Susan G. MacKay, Mohammed Bakir, Thomas J. Meyer, and Richard W. Linton, *X-ray Photoelectron Spectroscopy Sputter Depth Profile Analysis of Spatially Controlled Microstructures in Conductive Polymer Films* submitted July 27, 1990 and published January, 1991 in Analytical Chemistry, Volume 63, pgs. 60-65.
18. Fawcett, Steven C., and Dow, Thomas A., *Development of a Model for Precision Contour Grinding of Brittle Materials*, Precision Engineering Journal, accepted for publication March, 1991.

19. Hunter, J.L., Jr., and Russell, Phillip E., *Manipulation of Nanometer Scale Liquid Crystalline Fibrilles Using the Atomic Force Microscope* submitted to 38th Annual AVS Symposium and Topical Conference, May, 1991.
20. Shedd, Gordon M., and Russell, Phillip E., *The Effects of Low-Energy Ion Impacts on Graphite Observed by Scanning Tunneling Microscopy* published in *Journal of Vacuum Science and Technology A*, Vol. 9, No. 3, May/June 1991, p. 1261..
21. Ximen, Hongyu, I.H. Musselman, D. Bachelor, and P.E. Russell, *STM and AFM Tip Structures Microfabricated Using Focused Ion and Electron Beam Techniques*, abstract submitted to 38th Annual; AVS Symposium & Topical Conference, 05/15/91.
22. Ximen, Hongyu and Russell, Phillip E., *Atomic Force Microscopy Using Beam Fabricated Microtips*, submitted to *Applied Physics Letters*, 4/30/91.
23. Dow, Thomas A., Fawcett, Steven C., and Scattergood, Ronald O., *Ductile Regime Grinding of Brittle Materials*, abstract presented at National Institute of Standards & Technology, April 17, 1991.
24. Ro, Paul I., *Nonlinear Micro-Dynamic Behavior of a Ball-Screw Hydrostatic Slide System*, submitted to *Precision Engineering Journal*, February, 1991.
25. Musselman, I.H. and Russell, P.E., *Scanning Tunneling Microscopy and Atomic Force Microscopy of Fibrillar Structures in Liquid Crystalline Polymers*, submitted to the 1991 Meeting of the Electron Microscopy Society of America/Microbeam Analysis Society, March 1991.
26. Drescher, J.D. and Dow, T.A., *Machining Forces in Diamond Turning of Plated Copper and Unplated Substrates*, presented at American Society for Precision Engineering Spring Topical Meeting, April 15-18, 1991 in Tucson, AZ.
27. Bifano, T.G. and Fawcett, S.C., *Specific Grinding Energy as an In-Process Control Variable for Ductile-Regime Grinding*, *Precision Engineering Journal*, accepted for publication March 1991.
28. Miller, Michele H., Dow, Thomas A. and Falter, Peter J., *Application of a Fast Tool Servo for Diamond Turning of Non-Rotationally Symmetric Optical Surfaces* submitted to *Precision Engineering*, September, 1990. Published in *Precision Engineering: Journal of the American Society for Precision Engineering*, April, 1991.

29. Hiatt, Gary D. and Strenkowski, John S., *A Technique for Predicting the Ductile Regime in Single Point Diamond Turning* presented at the Winter Annual Meeting of the American Society of Mechanical Engineering Symposium, November 27, 1990, Dallas, TX.

THESIS/DISSERTATIONS

- 1.. Shedd, Gordan Michael, *Scanning Tunneling Microscopy Studies of Clusters and Materials Modification at the Nanometer Scale*, PhD Thesis, North Carolina State University, March, 1991.
2. Miller, Michelle, *Design of Three Axis Diamond Turning Machine Controller*, MS Thesis, North Carolina State University, May, 1991.
3. Sparks, Ronald, *Micro-Raman Investigation of Residual Stresses in Machined Semiconductors*, PhD Thesis, North Carolina State University, May, 1991.
4. Fawcett, Steven C., *Development and Implementation of a Grinding Technique for Precision Finishing of Brittle Materials*, PhD Thesis, North Carolina State University, July, 1991.
5. Abler, Jeffrey A., *Control of Precision Slide Motion for Vibration Reduction in Diamond Turning*, Master Thesis, North Carolina State University, December, 1991.
6. Hubbel, Peter I., *Modeling and Control of Nonlinear Microdynamics of a Ball-Screw Drive For Nanometer Motion*, North Carolina State University, December, 1991.
7. Tidwell, Michael, *Ductile Regime Machining of Germanium: Development of New Experimental and Analytical Analysis Methods*, North Carolina State University, December, 1991.